PERSPECTIVE
https://doi.org/10.1038/s41588-021-00885-0

nature genetics



# Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank

Joseph D. Szustakowski[1], Suganthi Balasubramanian[2], Erika Kvikstad[1], Shareef Khalid[2], Paola G. Bronson[3], Ariella Sasson[1], Emily Wong[4], Daren Liu[2], J. Wade Davis[5], Carolina Haefliger[6], A. Katrina Loomis[7], Rajesh Mikkilineni[4], Hyun Ji Noh[5], Samir Wadhawan[1], Xiaodong Bai[2], Alicia Hawes[2], Olga Krasheninina[2], Ricardo Ulloa[2], Alex E. Lopez[2], Erin N. Smith[4], Jeffrey F. Waring[5], Christopher D. Whelan[3], Ellen A. Tsai[3], John D. Overton[2], William J. Salerno[2], Howard Jacob[5], Sandor Szalma[4], Heiko Runz[3], Gregory Hinkle[8], Paul Nioi[8], Slavé Petrovski[6], Melissa R. Miller[7], Aris Baras[2], Lyndon J. Mitnaul[2], Jeffrey G. Reid[2]✉ and UKB-ESC Research Team*

**The UK Biobank Exome Sequencing Consortium (UKB-ESC) is a private–public partnership between the UK Biobank (UKB) and eight biopharmaceutical companies that will complete the sequencing of exomes for all ~500,000 UKB participants. Here, we describe the early results from ~200,000 UKB participants and the features of this project that enabled its success. The biopharmaceutical industry has increasingly used human genetics to improve success in drug discovery. Recognizing the need for large-scale human genetics data, as well as the unique value of the data access and contribution terms of the UKB, the UKB-ESC was formed. As a result, exome data from 200,643 UKB enrollees are now available. These data include ~10 million exonic variants—a rich resource of rare coding variation that is particularly valuable for drug discovery. The UKB-ESC precompetitive collaboration has further strengthened academic and industry ties and has provided teams with an opportunity to interact with and learn from the wider research community.**

Developing novel therapies to address unmet medical needs is a resource-intensive challenge characterized by high attrition rates[1]. While biopharmaceutical companies have recently improved overall research and development productivity and success rates for drug candidates in late-stage clinical development, the probability of a drug candidate proceeding from phase 1 clinical trials through approval and launch remains near 10%. Most clinical failures (~75%) are attributable to safety concerns or a lack of efficacy[2].

**Why are drug developers interested in human genetics?** The potential for human genetics to increase the likelihood of successful drug discovery has long been recognized, albeit largely in anecdotal form. Anti-*PCSK9* cholesterol-lowering drugs are a highly publicized example wherein human genetic evidence for a target contributed to technical and regulatory success, providing rationale for further investment[3]. Human genetics may also identify potential liability phenotypes associated with a target. For example, it is plausible that gastrointestinal adverse events observed in clinical trials of *DGAT1* inhibitors[4,5] could have been predicted based on the causal link between rare, highly penetrant *DGAT1* variants and congenital diarrheal disorder[6]. Genetics has also proven its value in bringing more precision to drug development, particularly in the context of large trials. Stratification of patients to enrich for signal has shown success in *PCSK9* inhibitors[7,8], providing the ability to select genetically defined patient populations.

Recently, several studies systematically characterized the role of human genetics in drug discovery[9–11]. These retrospective analyses of drug development successes and failures demonstrate that drug–target pairs with human genetic evidence are at least twice as likely to reach approval as those without. Moreover, there is evidence to suggest that targets with clear causal relationships to disease exhibit even higher success rates[10]. Furthermore, genetic association with a non-relevant phenotype increases the likelihood of corresponding adverse events[12].

Building on these and similar experiences, several frameworks for the systematic evaluation of genetically motivated targets have emerged. For example, the allelic series model leverages the existence of multiple, independent genetic alterations in a gene of interest to assess its potential as a drug target[13]. Genetic dose–response curves can be used to investigate the relationship between variants in the allelic series and phenotypes of interest to assess the potential for a tractable therapeutic window[14]. In addition, several promising drug development candidate targets with robust genetic evidence were identified based on associations between disease phenotypes, biomarker endo-phenotypes and functionally consequential genetic variants. Recent examples include *HSD17B13* for chronic liver disease[15], *TYK2* for multiple autoimmune disorders[16–18],

[1]Bristol Myers Squibb, Princeton, NJ, USA. [2]Regeneron Pharmaceuticals, Tarrytown, NY, USA. [3]Biogen, Cambridge, MA, USA. [4]Takeda California, San Diego, CA, USA. [5]Abbvie, North Chicago, IL, USA. [6]AstraZeneca Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, Cambridge, UK. [7]Pfizer, Cambridge, MA, USA. [8]Alnylam Pharmaceuticals, Cambridge, MA, USA. *A list of authors appear at the end of the paper.
✉e-mail: jeffrey.reid@regeneron.com

*NRXN1* for neuropsychiatric disease[19] and *ASGR1* for cardiovascular disease[20].

Large-scale biobanks have emerged to catalyze biomedical research, in part due to widespread adoption of electronic health records and the maturation of experimental and computational platforms capable of cost-effective population-scale sequencing and analysis. These advances create a unique opportunity for the scientific community to build on these foundations and accelerate the use of human genetics to inform drug discovery. Innovative public–private partnerships in the precompetitive space have demonstrated value for furthering this acceleration[21]. Here, we describe one such effort—the UK Biobank Exome Sequencing Consortium (UKB-ESC)—focused on generating whole-exome sequencing (WES) data for all participants (~500,000) in the UK Biobank (UKB).

**Why are drug developers interested in the UKB?** Nominating, evaluating and prioritizing potential drug targets based on human genetics is data intensive. The most actionable novel insights are likely to come from rare or private functional genetic variants leading to highly penetrant gains or losses of protein function[22]. These variants are most effectively discovered through direct sequencing of large populations or smaller populations likely to be enriched for variants of interest based on genetic architecture or the prevalence of specific traits. Human genetic data alone, however, are insufficient. Comprehensive, consistent, longitudinal phenotypic data that can be linked to the genetic data are needed to explore the breadth of biological consequences of genetic variation. Essential phenotypes may include accurate disease diagnosis, molecular and cellular biomarkers, treatment and outcome information, imaging endpoints, self-reported conditions and environmental and lifestyle data.

Beyond target identification, access to large-scale human genetic data linked to phenotypes enables a variety of other key opportunities for drug development efforts. In addition to direct human genetic discovery, these cohorts are incredibly valuable for instant replication of insights from other human genetics studies, disentangling causal relationships using methods such as Mendelian randomization, prediction of potential side effects, recall sub-studies, rapid validation of targets proposed by other means, and unbiased study of the natural history of rare monogenic diseases of interest to drug developers. Fifteen years ago, the Wellcome Trust Case Control Consortium shifted the field from candidate gene studies to genome-wide association studies[23]. Now, large biobanks are moving us even further in the unbiased spectrum by recruiting population-based cohorts without regard to disease status. The disease-agnostic approach and collection of a wide variety of data and specimens allow an immense number of hypotheses to be tested. The UKB is an exemplar of such data—a unique resource that provides a rich substrate for a broad spectrum of biomedical research[24].

The UKB design allows for recontact of participants to enroll them in new scientific research projects. This mechanism is highly valuable to the scientific community (including drug developers), as it allows researchers to engage with cohorts of specific, well-characterized participants, to address emerging scientific questions. As an example, the UKB recently launched a serology study to track the extent of infection of severe acute respiratory syndrome coronavirus 2.

The objective of the UKB-ESC is a comprehensive assessment of the protein-coding genetic variation in the half-million UKB participants. As the exome is enriched for variants that are most readily interpretable and actionable, this consortium has prioritized deep sequencing of the protein-coding portions of the genome. This is the highest-value assay to have been added to the UKB genotype resource, pairing the rarest coding variation with the common and rare variation captured via chip genotyping and imputation. The speed of WES is enabling a rapid acceleration of actionable scientific discoveries. While WES is the assay of choice for Mendelian disease discovery[25] and has provided important novel target discoveries[15], we are also enthusiastic for the arrival of the UKB whole-genome sequencing data[26] and recognize the value of continued investment in growing the UKB resource.

The enhancement of the UKB with WES data is not without limitations. It is well documented that human genetics studies are highly skewed towards populations of European ancestry[27,28]. Table 1 describes the clinical and demographic characteristics of the first 50,000 sequenced individuals (50,000 WES cohort), the current ~200,000 sequenced individuals (200,000 WES cohort) and the total of 500,000 participants (full UKB cohort), including information on self-reported ethnic background, which was captured in data field 21000. While the United Kingdom's population includes individuals from diverse backgrounds, its largest ethnic group is described as White British. In addition, the full UKB cohort is not representative of the UK population as a whole when considering ethnicity, age, sex, general health status and other factors[29]. Researchers need to be mindful that these data include a fraction of all human genetic variation and that signals derived from the UKB may not generalize to other populations.

**Lessons from drug discovery collaboration in genomics.** Large-scale collaborative projects have driven transformative scientific discoveries, particularly in genetics and genomics[24,30–35] (see also https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). Such collaborations provide a unique opportunity to unify scientific communities by enabling a diversity of voices and perspectives to produce insights that would be inaccessible to smaller, more homogeneous groups. Historically, large-scale collaborative life science projects have typically been funded by governments and non-profit organizations to engage academic groups. More recently, government agencies, non-profit organizations and academic institutions have sought industry partners for large-scale scientific projects (https://www.nih.gov/research-training/accelerating-medicines-partnership-amp, https://www.alzdiscovery.org/). The UKB-ESC seeks to further strengthen the ties between academia and industry through a precompetitive collaboration. This model is enabled by a project structure that engages all parties as scientific contributors and incentivizes industry investment in a sharable data resource. We expect that the output of this partnership will catalyze novel scientific discoveries, accelerate the development of new therapies and ultimately improve patient outcomes. As both the co-funders of the UKB WES data and the scientific teams who are analyzing them, we have found the following principles essential to the success of this effort.

First and foremost, the UKB-ESC is uniquely enabled by the UKB open data access policy. The biomedical research community has already made great use of UKB data, as evidenced by the rapidly growing body of scientific literature (https://www.ukbiobank.ac.uk/enable-your-research/publications). The UKB data access model is valuable in how it enables researchers to openly publish and commercialize results from and derivatives of the UKB data and incentivizes contributions back to the resource. This cycle enables academia and industry to be engaged and benefit from a body of work larger than any one entity can achieve alone.

Second, the UKB-ESC scale and scope will enable unique, valuable scientific discoveries. The search for actionable genetic variants (for example, rare with profound functional consequences or common with informative phenotypic associations) depends on the scale of available data, in terms of both sample size and phenotypic characterization. Building large, diverse and deeply characterized cohorts is an enormous multidisciplinary undertaking. Such work is particularly challenging in high-value phenotyping wherein different modalities (for example, imaging, proteomics and electronic medical record data extraction) require diverse expertise, and the

**Table 1 | Demographics and clinical characteristics of the publicly released 50,000 WES, 200,000 WES and full UKB cohorts**

| | 50,000 WES | | 200,000 WES | | Full UKB | |
|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % |
| Number of participants | 49,898 | | 200,633 | | 502,504 | |
| Female | 27,207 | 54.5 | 110,479 | 55.1 | 273,378 | 54.4 |
| Median age at assessment (years) | 58 | | 58 | | 58 | |
| Median body mass index (kg m$^{-2}$) | 26.6 | | 26.6 | | 26.7 | |
| Number of imaged participants | 14,174 | 28.4 | 23,818 | 11.9 | 44,703 | 8.9 |
| Number of current or past smokers | 22,327 | 44.7 | 89,391 | 44.6 | 226,901 | 45.2 |
| Median Townsend deprivation index score | −2.0 | | −2.2 | | −2.1 | |
| Inpatient ICD-10 codes per patient | 8 | | 8 | | 8 | |
| Patients with ≥1 ICD-10 diagnosis | 43,345 | 86.9 | 164,452 | 82.0 | 410,310 | 81.7 |
| **Self-reported ethnic background** | | | | | | |
| White | 46,536 | 93.3 | 188,027 | 93.7 | 472,114 | 94.0 |
| British | 43,333 | 86.8 | 176,037 | 87.7 | 442,575 | 88.1 |
| Irish | 1,497 | 3.00 | 5,413 | 2.70 | 13,207 | 2.63 |
| Any other white background | 1,706 | 3.42 | 6,577 | 3.28 | 16,332 | 3.25 |
| Mixed | 369 | 0.74 | 1,289 | 0.64 | 2,909 | 0.58 |
| White and Black Caribbean | 96 | 0.19 | 294 | 0.15 | 620 | 0.12 |
| White and Black African | 48 | 0.10 | 174 | 0.09 | 425 | 0.08 |
| White and Asian | 106 | 0.21 | 370 | 0.18 | 831 | 0.17 |
| Any other mixed background | 119 | 0.24 | 451 | 0.22 | 1,033 | 0.21 |
| Asian or Asian British | 1,066 | 2.14 | 4,247 | 2.12 | 9,839 | 1.96 |
| Indian | 708 | 1.42 | 2,622 | 1.31 | 5,951 | 1.18 |
| Pakistani | 139 | 0.28 | 776 | 0.39 | 1,837 | 0.37 |
| Bangladeshi | 17 | 0.03 | 89 | 0.04 | 236 | 0.05 |
| Any other Asian background | 202 | 0.40 | 760 | 0.38 | 1,815 | 0.36 |
| Black or Black British | 1,005 | 2.01 | 3,220 | 1.60 | 8,034 | 1.60 |
| Caribbean | 658 | 1.32 | 1,931 | 0.96 | 4,517 | 0.90 |
| African | 337 | 0.68 | 1,241 | 0.62 | 3,394 | 0.68 |
| Any other Black background | 10 | 0.02 | 47 | 0.02 | 123 | 0.02 |
| Chinese | 173 | 0.35 | 643 | 0.32 | 1,574 | 0.31 |
| Other ethnic group | 505 | 1.01 | 1,929 | 0.96 | 4,558 | 0.91 |
| Do not know | 14 | 0.03 | 75 | 0.04 | 217 | 0.04 |
| Prefer not to answer | 177 | 0.35 | 719 | 0.36 | 1,661 | 0.33 |
| **Enhanced measures** | | | | | | |
| Hearing test | 40,761 | 81.7 | 84,484 | 42.1 | 171,826 | 34.2 |
| Pulse rate | 40,495 | 81.2 | 83,974 | 41.9 | 170,743 | 34.0 |
| Visual acuity measured | 39,551 | 79.3 | 65,248 | 32.5 | 117,672 | 23.4 |
| IOP measured (left) | 37,889 | 75.9 | 62,213 | 31.0 | 111,926 | 22.3 |
| Autorefraction | 39,520 | 79.2 | 64,807 | 32.3 | 116,365 | 23.2 |
| Retinal OCT | 38,371 | 76.9 | 51,286 | 25.6 | 78,888 | 15.7 |
| ECG at rest | 12,331 | 24.7 | 15,953 | 8.0 | 24,923 | 5.0 |
| Cognitive function | 17,340 | 34.8 | 53,039 | 26.4 | 123,614 | 24.6 |
| Digestive health | 23,922 | 47.9 | 74,417 | 37.1 | 174,760 | 34.8 |
| Physical activity measurement | 15,002 | 30.1 | 44,885 | 22.4 | 103,684 | 20.6 |
| **Cardiometabolic phenotypes** | | | | | | |
| Coronary disease | 4,356 | 8.7 | 16,224 | 8.1 | 45,450 | 9.0 |
| Heart failure | 707 | 1.4 | 3,009 | 1.5 | 8,939 | 1.8 |
| Type 2 diabetes | 2,817 | 5.6 | 11,150 | 5.6 | 29,266 | 5.8 |

Continued

**Table 1 | Demographics and clinical characteristics of the publicly released 50,000 WES, 200,000 WES and full UKB cohorts (Continued)**

| | 50,000 WES | | 200,000 WES | | Full UKB | |
|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % |
| **Respiratory and immunological phenotypes** | | | | | | |
| Asthma | 8,471 | 17.0 | 28,491 | 14.2 | 70,969 | 14.1 |
| COPD | 1,481 | 3.0 | 5,135 | 2.6 | 14,688 | 2.9 |
| Rheumatoid arthritis | 951 | 1.9 | 3,802 | 1.9 | 9,685 | 1.9 |
| Inflammatory bowel disease | 650 | 1.3 | 2,616 | 1.3 | 6,851 | 1.4 |
| **Neurodegenerative phenotypes** | | | | | | |
| Alzheimer's disease | 47 | 0.09 | 352 | 0.18 | 933 | 0.19 |
| Parkinson's disease | 131 | 0.26 | 721 | 0.36 | 1,938 | 0.39 |
| Multiple sclerosis | 163 | 0.33 | 684 | 0.34 | 1,766 | 0.35 |
| Myasthenia gravis | 23 | 0.05 | 129 | 0.06 | 300 | 0.06 |

Phenotypes were defined using a combination of International Classification of Diseases, Tenth Revision (ICD-10) codes together with information from a self-reported verbal questionnaire. Values are expressed as medians. The data correspond to UKB phenotypic release basket UKBB_41065 (March 2020). Participants withdrawn from the UKB as of September 2020 were excluded. Data on ethnic background were taken from UKB field 21000. COPD, chronic obstructive pulmonary disease; ECG, electrocardiogram; IOP, intraocular pressure; OCT, optical coherence tomography.

**Table 2 | Summary statistics for variants in 200,643 publicly released exomes**

| | Variants in WES | | Median per participant | | | |
|---|---|---|---|---|---|---|
| | Number of variants | Number of variants with MAF < 1% | Number of variants | IQR | Number of variants with MAF < 1% | IQR |
| Total | 16,846,188 | 16,709,690 | 44,719 | 567 | 3,050 | 107 |
| Targeted regions | 8,457,134 | 8,395,913 | 21,380 | 240 | 1,466 | 50 |
| **Variant type** | | | | | | |
| SNVs | 8,086,176 | 8,026,702 | 20,817 | 234 | 1,403 | 49 |
| Indels | 370,958 | 369,211 | 563 | 23 | 64 | 8 |
| Multi-allelic sites | 1,596,984 | 1,585,412 | 3768 | 68 | 223 | 19 |
| **Functional prediction** | | | | | | |
| Synonymous | 2,139,318 | 2,115,702 | 8,586 | 114 | 457 | 24 |
| Missense | 4,549,694 | 4,526,159 | 7,724 | 115 | 677 | 32 |
| LOF (any transcript) | 453,733 | 453,007 | 202 | 15 | 31 | 7 |
| LOF (strict) | 381,717 | 381,232 | 142 | 12 | 22 | 6 |

Counts of autosomal variants observed across all individuals by type and functional class, for all variants and for those with a MAF of <1%, are shown. The number of bases targeted for capture by the exome capture reagent was $n = 38{,}997{,}831$. Median counts and interquartile ranges (IQRs) per individual for all variants and for those with a MAF of <1% are also shown. Counts were restricted to WES targeted regions. Variant annotation details are included in the Methods.

data often require extensive curation and processing before research use. The UKB excels at such tasks within a user-friendly framework attractive to both academic and industrial researchers, creating a roadmap for other projects with similar aspirations.

Third, the UKB's data access and contribution terms invite pre-competitive collaboration by industry partners. The generation of sequencing data at this scale requires a substantial investment of time, personnel and financial resources[36]. The UKB policy of providing a window of data exclusivity for data generators (also a common feature of large-scale academic collaborations) was essential to the business case for an investment of this scale. In addition to the scientific drivers described above, participation in a consortium and the exclusivity period both provide tangible benefits. While the competitive commercial interests of companies may seem to preclude cooperation, an appropriately managed collaboration mitigates individual risk to each partner and provides value larger than the sum of the individual investments. Moreover, the finite window of exclusivity is an incentive for the partner companies to focus their initial efforts on projects that are likely to have near-term impact on their pipelines.

Finally, the value of engagement in a large, precompetitive industry collaborative project provides additional value to the participating institutions. Building expertise and functional excellence in important new areas is a key benefit to participating institutions. In our experience, drug discovery teams typically have deep experience in areas such as disease biology, chemistry and translational research. In contrast, the inclusion of human genetics expertise varies between companies and therapeutic areas and is often rate limited by the availability of relevant datasets. Participation in the UKB-ESC provided member companies with a unique opportunity to build out or enhance the expertise needed to conduct population-scale genomic medicine research without each incurring the full overhead of creating the underlying resource. The resulting combination of scientific expertise and a critical mass of data serves as a multiplier for extant research programs, complementing existing genomics data, such as targeted clinical sequencing of probands and families, and providing a roadmap for at-scale genomics in therapeutic areas that remain under-represented in genetic data. We anticipate a virtuous cycle between the increasing availability of large, well-annotated genetics resources and increased industry investment in genetics.

In summary, we have identified the following key features of an effective precompetitive industry collaboration: (1) build a large dataset with rich phenotypic characterization, noting that participant recontact is highly valued for answering new questions that emerge from the data; (2) provide researchers with the opportunity to derive both academic and commercial value from the data in an unencumbered way; (3) provide some exclusivity/first-mover advantage to build a compelling business justification for participation and financing of the project; (4) enable data access providing insight generation for key internal therapeutic-focused scientists and research and development stakeholders within the collaborating institutions; and (5) provide opportunities for constructive engagement with academic partners, as well as low-friction data-sharing terms and platforms to enable the broadest possible suite of research projects. We are hopeful that others will follow the UKB model and build substantial data collections designed to engage a wide variety of stakeholders.

## Results

The release of the first 200,633 sequenced exomes represents a milestone in the availability of large-scale genomics data. This release is inclusive of the first 50,000 UKB samples[37], the sequencing and initial release of which were funded through a separate mechanism independent of the UKB-ESC. Here, we provide an overview of the phenotypes and genotypes available as part of this resource. Table 1 includes a summary of the clinical characteristics of the 50,000 WES, 200,000 WES and full UKB cohorts. Definitions of the UKB phenotypes are provided in ref. [37]. Table 2 summarizes the variants observed in the first 200,000 exomes sequenced from the UKB (that is, the 200,000 WES cohort). The targeted region covered 38,997,831 bases and coverage exceeded 20× on 95.6% of the sites on average. Approximately 10 million variants were observed within the targeted regions. These include 8,086,176 single-nucleotide variants (SNVs), 370,958 indels and 1,596,984 multi-allelic variants. Of the ~8 million SNVs observed, 84.5% are coding variants and include 2,139,318 (25.3%) synonymous variants, 4,549,694 (53.8%) missense variants and 453,733 (5.4%) predicted loss-of-function (LOF) variants (initiation codon loss, premature stop codons, stop codon loss or splicing and frameshift variants) affecting at least one coding transcript. Analysis of allele frequencies revealed 98% of synonymous, 99% of missense and 99% of LOF (at least one transcript) variants with a minor allele frequency (MAF) of <1%. On a per-individual basis, we observed a median number of 8,586 synonymous, 7,724 missense and 202 LOF variants. Restricting our analysis to variants that affect canonical transcripts, we observed a median of 142 LOFs per individual. The numbers are largely similar to those in published exome studies[37,38].

In addition, we also analyzed the increase in the number of genes with heterozygous and homozygous LOFs with the increase in the number of sequenced samples. Previously, 17,718 genes with heterozygous LOF variants were observed in the 50,000 WES data[37]; therefore, is it not surprising that we only observed a modest increase in that number (18,045 genes) in the 200,000 WES data. A total of 789 genes with at least one homozygous LOF variant were reported in the 50,000 WES data. The number of genes with homozygous LOFs still appears to be increasing, with a projection of 1,981 genes with at least one homozygous LOF variant in the 500,000 participants of the full UKB cohort (1,492 genes with at least one homozygous LOF seen in the 200,000 WES cohort; Table 3). The tolerance of homozygous LOF variants is of particular interest to target development programs, although characterization of homozygous LOFs in all human genes will not be accomplished by scale alone[39]. Rather, the increasing number of genes tolerating homozygous LOFs identified in the UKB exomes can complement smaller study designs, such as ancestry-specific population sequencing and consanguineous cohorts.

**Table 3 | Observed numbers of heterozygous and homozygous carriers of LOF variants with an AAF of <1% in ~200,000 exomes, and projections for the numbers expected in 500,000 exomes**

| Minimum number of carriers with LOF (AAF < 1%) (*n*) | Observed in 189,698 individuals | Predicted in 500,000 individuals |
|---|---|---|
| Numbers of genes with at least *n* heterozygous carriers | | |
| 1 | 18,045 | 18,414 |
| 5 | 17,411 | 17,968 |
| 10 | 16,556 | 17,514 |
| 25 | 14,204 | 16,384 |
| 50 | 11,380 | 14,853 |
| 100 | 7,830 | 12,475 |
| 250 | 3,824 | 8,065 |
| 500 | 1,905 | 4,711 |
| 1,000 | 914 | 2,449 |
| 2,500 | 238 | 924 |
| 5,000 | 18 | 394 |
| 10,000 | 1 | 57 |
| Numbers of genes with at least *n* homozygous carriers | | |
| 1 | 1,492 | 1,981 |
| 5 | 528 | 954 |
| 10 | 193 | 627 |
| 25 | 20 | 199 |
| 50 | 2 | 42 |
| 100 | 0 | 5 |
| 250 | 0 | 0 |
| 500 | 0 | 0 |
| 1,000 | 0 | 0 |
| 2,500 | 0 | 0 |
| 5,000 | 0 | 0 |
| 10,000 | 0 | 0 |

The numbers of autosomal genes with at least *n* heterozygous and homozygous carriers of LOFs with an alternative allele frequency (AAF) of <1% and passing the quality control filters described in Table 2 are shown for 189,698 UKB participants of European ancestry (the 200,000 WES cohort). Predicted numbers, estimated based on the methodology described in ref. [37], are also shown for such genes in the 500,000 individuals of the full UKB cohort.

## Discussion

The UKB exists to enable scientific research with the ultimate aim of improving human health. The UKB-ESC is proud of its contribution to the scientific community, enthusiastic about making these data broadly available and excited to see what the future holds in terms of discoveries from and contributions to the UKB and UKB-ESC resources, particularly as additional insight is gained into functionally consequential variants that can meaningfully inform drug development. Data and methods developed by the UKB-ESC have already contributed to multiple publications[40–44].

With the UKB-ESC exome sequencing nearly complete, we hope the key features of this collaboration will be adopted as the preferred model for similar projects in the future. We expect that the value of the WES data will be enhanced by layering deeper and richer phenotypes, which can provide important insights into disease biology, characterize responses to marketed therapies and identify novel targets. Thoughtfully designed projects that maximize the

engagement of academia, non-profit organizations and industry will yield valuable data resources and scientific insights that accelerate drug discovery and personalized health care. As an example, the recently announced Pharma Proteomics Project will use a model similar to that of the UKB-ESC to generate high-dimensional proteomics data on approximately 53,000 participants[45]. The groundwork laid by these private–public partnerships will nucleate further partnerships by lowering entry costs: existing partners already have relationships established and new partners have example data and contractual frameworks to draw on to make the business case.

Developing new therapies to treat unmet medical needs is among the most important scientific challenges we face. Large-scale collaborations such as the UKB-ESC play an essential role by generating unique, accessible resources that can be used by a diverse community of researchers to address questions critical to advancing human health.

## Data availability

The UKB aims to encourage and provide the widest possible access to its data and samples for health-related research in the public interest performed by all bona fide researchers from the academic, charity, public and commercial sectors, both in the United Kingdom and internationally, without preferential access for any user. The UKB's publicly available Data Showcase (http://biobank.ndph.ox.ac.uk/showcase/) presents the univariate distributions and methods used for collection of all of the variables available for health-related research, enabling potential research users to explore which data are available and to plan research applications. All researchers who wish to access the resource must register with the UKB via its online access management system (https://bbams.ndph.ox.ac.uk/ams/). Once approved, researchers may apply (via the access management system) to access the resource for specific, well-defined research projects. At the time of publication, over 16,500 researchers were registered with the UKB and over 2,000 research applications were approved (see https://www.ukbiobank.ac.uk/enable-your-research/approved-research for a summary of research that is currently underway).

Received: 6 November 2020; Accepted: 13 May 2021;
Published online: 28 June 2021

## References

1. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
2. Dowden, H. & Munro, J.Trends in clinical success rates and therapeutic focus. *Nat. Rev. Drug Discov.* **18**, 495–496 (2019).
3. Furtado, R. H. M. & Giugliano, R. P.What lessons have we learned and what remains to be clarified for PCSK9 inhibitors? A review of FOURIER and ODYSSEY outcomes trials. *Cardiol. Ther.* **9**, 59–73 (2020).
4. Denison, H. et al. Proof of mechanism for the DGAT1 inhibitor AZD7687: results from a first-time-in-human single-dose study. *Diabetes Obes. Metab.* **15**, 136–143 (2013).
5. Meyers, C. D. et al. Effect of the DGAT1 inhibitor pradigastat on triglyceride and apoB48 levels in patients with familial chylomicronemia syndrome. *Lipids Health Dis.* **14**, 8 (2015).
6. Haas, J. T. et al. *DGAT1* mutation is linked to a congenital diarrheal disorder. *J. Clin. Invest.* **122**, 4680–4684 (2012).
7. Sabatine, M. S. et al. Evolocumab and clinical outcomes in patients with cardiovascular disease. *N. Engl. J. Med.* **376**, 1713–1722 (2017).
8. Schwartz, G. G. et al. Alirocumab and cardiovascular outcomes after acute coronary syndrome. *N. Engl. J. Med.* **379**, 2097–2107 (2018).
9. Cook, D. et al. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–431 (2014).
10. King, E. A., Wade Davis, J. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
11. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
12. Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P. & Ward, L. D. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat. Commun.* **10**, 1579 (2019).
13. McKusick, V. A.Phenotypic diversity of human diseases resulting from allelic series. *Am. J. Hum. Genet.* **25**, 446–456 (1973).
14. Plenge, R. M., Scolnick, E. M. & Altshuler, D.Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
15. Abul-Husn, N. S. et al. A protein-truncating *HSD17B13* variant and protection from chronic liver disease. *N. Engl. J. Med.* **378**, 1096–1106 (2018).
16. Dendrou, C. A. et al. Resolving *TYK2* locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* **363**, 363ra149 (2016).
17. Diogo, D. et al. *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 (2015).
18. Minegishi, Y. et al. Human tyrosine kinase 2 deficiency reveals its requisite roles in multiple cytokine signals involved in innate and acquired immunity. *Immunity* **25**, 745–755 (2006).
19. Noh, H. J. et al. Integrating evolutionary and regulatory information with a multispecies approach implicates genes and pathways in obsessive-compulsive disorder. *Nat. Commun.* **8**, 774 (2017).
20. Nioi, P. et al. Variant *ASGR1* associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.* **374**, 2131–2141 (2016).
21. Dolgin, E. Massive NIH–industry project opens portals to target validation. *Nat. Rev. Drug Discov*. https://doi.org/10.1038/d41573-019-00033-8 (2019).
22. McCarthy, M. I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
23. Burton, P. R. et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
24. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
25. Yang, Y. et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
26. UK Biobank leads the way in genetics research to tackle chronic diseases. *UK Biobank* https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-leads-the-way-in-genetics-research-to-tackle-chronic-diseases-1 (2019).
27. Diversity matters. *Nat. Rev. Genet.* **20**, 495 (2019).
28. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S.Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).
29. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
30. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
31. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
32. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
33. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
34. Methé, B. A. et al. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
35. Holden, A. L., Contreras, J. L., John, S. & Nelson, M. R.The international serious adverse events consortium. *Nat. Rev. Drug Discov.* **13**, 795–796 (2014).
36. Regeneron announces major collaboration to exome sequence UK Biobank genetic data more quickly. *UK Biobank* https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/regeneron-announces-major-collaboration-to-exome-sequence-uk-biobank-genetic-data-more-quickly (2018).
37. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
38. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
39. Minikel, E. V. et al. Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459–464 (2020).
40. Liu, J. Z. et al. The burden of rare protein-truncating genetic variants on human lifespan. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.02.129908 (2020).
41. Povysil, G. et al. Assessing the role of rare genetic variation in patients with heart failure. *J. Am. Med. Assoc. Cardiol.* **6**, 379–386 (2021).
42. Carss, K. J. et al. Spontaneous coronary artery dissection: insights on rare genetic variation from genome sequencing. *Circ. Genom. Precis. Med.* **13**, e003030 (2020).
43. Dhindsa, R. et al. Identification of a novel missense variant in *SPDL1* associated with idiopathic pulmonary fibrosis. *Commun. Biol.* **4**, 392 (2021).
44. Cameron-Christie, S. et al. A broad exome study of the genetic architecture of asthma reveals novel patient subgroups. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.10.419663 (2020).

45. UK Biobank launches one of the largest scientific studies measuring circulating proteins, to better understand the link between genetics and human disease. *UK Biobank* https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/uk-biobank-launches-one-of-the-largest-scientific-studies (2020).

## Acknowledgements

We thank everyone who made this work possible, including the UKB team, their funders and the dedicated professionals from the member institutions who contributed to and supported this work. We are especially grateful to the UKB participants who generously volunteered to take part in this research. This research has been conducted using the UKB resource under application number 26041.

## Author contributions

J.D.S., P.G.B., E.W., J.W.D., C.H., A.K.L., R.M., H.J.N., S.W., E.N.S., J.F.W., C.D.W., E.A.T., J.D.O., W.J.S., H.J., S.S., H.R., G.H., P.N., S.P., M.R.M., A.B., L.J.M. and J.G.R. jointly supervised the research. J.D.S., W.J.S., A.B. and J.G.R. conceived of and designed the experiments. A.E.L. and J.D.O. performed the experiments. J.D.S., S.B., A.S., S.K., E.K., D.L., X.B., A.H., O.K. and W.J.S. analyzed the data. X.B., A.H., O.K., R.U., W.J.S. and J.G.R. contributed reagents, materials and/or analysis tools. J.D.S., S.B., A.S., P.G.B., E.K., W.J.S., S.S., H.R., P.N., S.P., M.R.M. and J.G.R. wrote the paper.

## Competing interests

J.D.S. is employed by and owns stocks in Bristol Myers Squibb. C.D.W. and H.R. are employees of Biogen and hold stock options at Biogen. E.N.S. is an employee of Takeda California. S.S. is employed by Takeda California and is a shareholder of Takeda and Johnson & Johnson. H.J.N. and J.W.D. are employed by AbbVie and may own AbbVie stock and/or options. A.K.L. is an employee of Pfizer. M.R.M. is an employee and stockholder of Pfizer. P.N. is an employee of and stockholder in Alnylam Pharmaceuticals. G.H. is a paid consultant to 54gene—a Nigerian genomics company. C.H. and S.P. are employees and stockholders of AstraZeneca. W.J.S., R.U., A.H., O.K., S.K., D.L., X.B., A.E.L., S.B., A.B., J.D.O. and L.J.M. are full-time employees of the Regeneron Genetics Center of Regeneron Pharmaceuticals and receive stock options and restricted stock units as compensation. J.G.R. is a full-time employee of the Regeneron Genetics Center of Regeneron Pharmaceuticals and receives stock options and restricted stock units as compensation. J.G.R. also provides (unpaid) advice, support and guidance to the UKB as a member of the UKB International Scientific Advisory Board. The other authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00885-0.

**Correspondence** should be addressed to J.G.R.

**Peer review information** *Nature Genetics* thanks Amalio Telenti, Matthew Nelson and Kaja Wasik for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



## UKB-ESC Research Team

**Oleg Moiseyenko[1], Carlos Rios[1], Saurabh Saha[1], Goncalo Abecasis[2], Nilanjana Banerjee[2], Christina Beechert[2], Boris Boutkov[2], Michael Cantor[2], Giovanni Coppola[2], Aris Economides[2], Gisu Eom[2], Caitlin Forsythe[2], Erin D. Fuller[2], Zhenhua Gu[2], Lukas Habegger[2], Marcus B. Jones[2], Rouel Lanche[2], Michael Lattari[2], Michelle LeBlanc[2], Dadong Li[2], Luca A. Lotta[2], Kia Manoochehri[2], Adam J. Mansfield[2], Evan K. Maxwell[2], Jason Mighty[2], Mrunali Nafde[2], Sean O'Keeffe[2], Max Orelus[2], Maria Sotiropoulos Padilla[2], Razvan Panea[2], Tommy Polanco[2], Manasi Pradhan[2], Ayesha Rasool[2], Thomas D. Schleicher[2], Deepika Sharma[2], Alan Shuldiner[2], Jeffrey C. Staples[2], Cristopher V. Van Hout[2], Louis Widom[2], Sarah E. Wolf[2], Sally John[3], Chia-Yen Chen[3], David Sexton[3], Varant Kupelian[3], Eric Marshall[3], Timothy Swan[3], Susan Eaton[3], Jimmy Z. Liu[3], Stephanie Loomis[3], Megan Jensen[3], Saranya Duraisamy[3], Jason Tetrault[4], David Merberg[4], Sunita Badola[4], Mark Reppell[5], Jason Grundstad[5], Xiuwen Zheng[5], Aimee M. Deaton[8], Margaret M. Parker[8], Lucas D. Ward[8], Alexander O. Flynn-Carroll[8], Caroline Austin[6], Ruth March[6], Menelas N. Pangalos[6], Adam Platt[6], Mike Snowden[6], Athena Matakidou[6], Sebastian Wasilewski[6], Quanli Wang[6], Sri Deevi[6], Keren Carss[6], Katherine Smith[6], Morten Sogaard[7], Xinli Hu[7], Xing Chen[7] and Zhan Ye[7]**