The impact of non-additive genetic associations on age-related complex diseases

Marta Guindo-Martínez, Ramon Amela, Silvia Bonàs-Guarch, Montserrat Puiggròs, Cecilia Salvoro, Irene Miguel-Escalada, Caitlin E. Carey, Joanne B. Cole, Sina Rüeger, Elizabeth Atkinson, Aaron Leong, Friman Sanchez, Cristian Ramon-Cortes, Jorge Ejarque, Duncan S. Palmer, Mitja Kurki, FinnGen Consortium, Krishna Aragam, Jose C. Florez, Rosa M. Badia, Josep M. Mercader, and David Torrents.

Genome-wide association studies (GWAS) are not fully comprehensive, as current strategies typically test only the additive model, exclude the X chromosome, and use only one reference panel for genotype imputation. We implement an extensive GWAS strategy, GUIDANCE, which improves genotype imputation by using multiple reference panels and includes the analysis of the X chromosome and non-additive models to test for association. We apply this methodology to 62,281 subjects across 22 age-related diseases and identify 94 genome-wide associated loci, including 26 previously unreported. Moreover, we observe that 27.7% of the 94 loci are missed if we use standard imputation strategies with a single reference panel, such as HRC, and only test the additive model. Among the new findings, we identify three novel low-frequency recessive variants with odds ratios larger than 4, which need at least a three-fold larger sample size to be detected under the additive model. This study highlights the benefits of applying innovative strategies to better uncover the genetic architecture of complex diseases.

https://doi.org/10.1038/s41467-021-21952-4

1 Barcelona Supercomputing Center (BSC), Barcelona, Spain. 2 Regulatory Genomics and Diabetes, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain. 3 CIBER de Diabetes y Enfermedades Metabólicas Asociadas, Madrid, Spain. 4 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 5 Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. 6 Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 7 Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. 8 Harvard Medical School, Boston, MA, USA. 9 Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children’s Hospital, Boston, MA, USA. 10 Institute for Molecular Medicine Finland, FIMM, HIIT, University of Helsinki, Helsinki, Finland. 11 Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 12 Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. 13 Cardiology Division, Massachusetts General Hospital, Boston, MA, USA. 14 Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. 15 Department of Medicine, Harvard Medical School, Boston, MA, USA. 16 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. 17 Present address: GENOMICS plc, Oxford, UK. 18 These authors contributed equally: Marta Guindo-Martínez, Ramon Amela. 19 These authors jointly supervised this work: Josep M. Mercader, David Torrents. 20 A full list of members and their affiliations appears in the Supplementary Information. 21 Email: mercader@broadinstitute.org; david.torrents@bsc.es
Genome-wide association studies (GWAS) have been successful in identifying thousands of associations between genetic variation and human complex diseases and traits. Nevertheless, for most complex diseases, only a small fraction of their genetic architecture is known, and a small amount of the estimated heritability is explained. Variants that individually have small contributions to the risk of disease, and/or are rare in the population, are often missed by the majority of GWAS even though they may contribute to the pathophysiology of complex diseases. Some of the current limitations of GWAS could be overcome by increasing sample sizes and by applying more comprehensive analytical methods with improved imputation strategies. Besides, whole-exome sequencing (WES) and whole-genome sequencing (WGS) datasets, such as the NHLBI Trans-Omics for Precision Medicine (TOPMed) Project, and the UK Biobank, are rapidly growing and expanding the range of variants to be tested in genetic association studies. However, together with the general tendency of increasing sample sizes, this imposes additional methodological and computational challenges. These can require scientists to restrict and simplify the analysis by limiting it to autosomal chromosomes, to a single reference panel for imputation, and to a single (additive) inheritance model for association testing, leaving a relevant fraction of the genetic architecture of the disease unexplored.

The genetic variants that modify the risk to develop a particular complex disease may contribute to the final phenotype through different functional mechanisms defined by a particular model of inheritance, which is further reflected in a characteristic distribution of affected alleles across patients and healthy individuals in GWAS. For example, the additive inheritance model, which is often the only genetic model tested, assumes that the risk of the disease is proportional to the number of risk alleles in an individual, i.e., that the effect of the heterozygous genotype is halfway between the two possible homozygous genotypes. However, some variants follow non-additive inheritance models, which include dominant, recessive, and heterodominant. The additive model is expected to capture a large fraction of the genetic risk for disease and can identify some variants that show recessive effects. The importance of evaluating non-additive inheritance models is well reported in the context of Mendelian diseases and occasionally for complex traits as well, such as the recessive effects of the FTO locus in obesity, and in or near ITGA12, TBC1D4, and CDKAL1 genes in type 2 diabetes, as well as the known non-additive effects of HLA haplotypes in autoimmune diseases and ulcerative colitis. The increasing ability to capture low-frequency variants using modern imputation reference panels and the need to uncover the still missing heritability estimated for most complex diseases, call for comprehensive association strategies that should include, among other improvements, the analysis of non-additive inheritance models.

In this work, and to fill this gap and to determine the prevalence and contribution of the different inheritance patterns involved in the genetic architecture of complex diseases, we design and implement a comprehensive strategy for genetic association analysis that combines imputation from multiple reference panels with association testing under five different inheritance models across multiple phenotypes. We apply this strategy to the Kaiser Permanente Research Program on Genes, Environment and Health: A Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort, which includes 62,281 subjects from European ancestry and 22 diseases. Finally, we release here both the summary statistics for all the models of inheritance as well as the complete methodology, provided to the community as an easy-to-use and standalone pipeline. This pipeline allows the analysis of existing and newly generated GWAS data with better efficiency and more comprehensive testing, improving the chances of variant discovery.

Results
In order to assess the potential benefits of applying more in-depth GWAS methodologies to available genetic datasets, and to investigate the relative contribution of different inheritance models to the risk to develop complex diseases, we have applied a global analysis strategy to the GERA cohort, an age-related disease-based cohort with an average age of 63, well-powered to study a broad range of clinically defined age-related conditions. By using this particular cohort, we expect to minimize a possible loss of power due to the misclassification of controls, as often happens in datasets with younger individuals that can include cases at pre-disease stages classified as controls.

Genotype imputation and association testing using multiple reference panels. After applying genetic quality control to the GERA cohort (see “Methods”), we retained 56,637 individuals with European ancestry for further downstream analysis (Supplementary Data 1). To cover the maximum number and type of genetic variants, we next applied an imputation strategy with four reference panels: the Genome of the Netherlands (GoNL)18,19, the UK10K Project20, the 1000 Genomes Project (1000G) phase 321, and Haplotype Reference Consortium (HRC)22, and imputed 11.2 M, 11.4 M, 13.1 M, and 11.7 M high-quality imputed variants (IMPUTE2 info score ≥ 0.7 and minor allele frequency [MAF] ≥ 0.001) with each panel, respectively. After combining the results of the four reference panels by choosing, for each variant, the panel that provided the highest imputation accuracy, we retained a total of 16,059,686 variants covering all the autosomes and the X chromosome (Fig. 1a). Using this strategy we imputed 2.6 M and 5.5 M high-quality, low-frequency (0.05 > MAF > 0.01) and rare variants (0.01 > MAF > 0.001), respectively, as well as 1.6 M indels. Note that as many as 684,393 common variants (MAF ≥ 0.05), 255,106 low-frequency, 1.7 M rare, and all indels (1.6 M) would be missed if only the HRC reference panel was used. This highlights the benefit of combining different reference panels for comprehensive association testing (Fig. 1b).

To evaluate our imputation strategy, we used sequenced data from UK10K, which includes 3781 sequenced genomes, to build an in silico array of 599,208 variants. We then imputed the in silico array using 1000G, GoNL and HRC as reference panels and compared the imputed genotype dosages with the sequenced genotype dosages (allelic dosage $R^2$, see “Methods” and Supplementary Fig. 1). Our results empirically demonstrate that our IMPUTE2-info threshold of 0.7 is a good cutoff for well-predicted genotypes (Supplementary Fig. 2). Moreover, the combination of the results from the different panels based on IMPUTE2-info values outperforms single reference panels in terms of % of variants with allelic dosage $R^2$ ≥ 0.5 in autosomes (Fig. 1c) and in the X chromosome (Fig. 1d) for both SNPs and indels (Supplementary Fig. 3).

To evaluate that our strategy is able to identify previously known loci that are well-powered in our cohort, we performed a power comparison under the additive model for type 2 diabetes and age-related macular degeneration, since both are well-defined diseases with available summary statistics. The power comparison between our additive results for type 2 diabetes and those of the DIAMANTE consortium showed that while we had an 80% power to find two variants with a p-value of 5.0 × 10⁻⁸ in
our dataset, we found four variants with genome-wide significance (Supplementary Data 2). For age-related macular degeneration, we had 80% power to find six variants with a *p*-value of $5.0 \times 10^{-8}$ of which we identified four of them with genome-wide significance. The remaining two variants were nominally significant in our study (Supplementary Data 3). These analyses suggest that our strategy is robust to find previously known variants given sufficient power.

Fig. 1 Graphical representation illustrating the benefits of combining the results from different reference panels. **a** Comparison of the number of variants after the imputation with four reference panels (info score ≥ 0.7), and combining them, colored according to MAF and variant type (SNP vs alternative forms of variation, such as indels). As shown in the bar plot, combining the results from the four reference panels increased the final set of variants for association testing when compared with the results for each of the panels alone (GoNL, UK10K, 1000G Phase 3, or HRC), especially in the low and rare frequency spectrum. For example, we covered up to 5.5 M rare variants (0.01 > MAF > 0.001) by combining panels, while only 2.3 M, 2.9 M, 3.2 M, and 3.8 M of rare variants were imputed independently with GoNL, UK10K, 1000G phase 3, and HRC, respectively. **b** Comparison of the contribution of each reference panel in the combined results. Each bar represents the number of variants that had the best imputation accuracy for a given reference panel. As shown in the figure, although the HRC panel showed overall higher imputation scores, as it provided around 10% of the final 16 M variants, the contribution of the other reference panels, primarily with non-SNP variants, was substantial. Indels seen in the bar plot for HRC correspond to genotyped indels. All variants with info score <0.7, MAF < 0.001, and HWE for controls *p* < 1.0 × 10^{-6} were filtered. **c** Percentage of high-quality imputed variants (IMPUTE2-info score ≥ 0.7) with an allelic dosage $R^2$ ≥ 0.5 between sequenced genotypes in UK10K samples vs variants imputed in the same UK10K samples using 1000G phase 3, GoNL, and HRC reference panels for the autosomes. The percentage of high-quality imputed variants with allelic dosage $R^2$ values (y axis) are represented across several MAF ranges (x-axis) for each of the reference panels and the combined panels imputed results. The combination of the three reference panels outperforms the single reference panels with 97.74% of variants with allelic dosage $R^2$ ≥ 0.5. **d** Percentage of variants in the X chromosome with an IMPUTE2-info score ≥ 0.7 and with an allelic dosage $R^2$ ≥ 0.5 for UK10K imputed genotypes across MAF ranges for 1000G phase 3, GoNL, and HRC reference panels and the combined results. The combination of the results from the three panels outperforms single reference panels with 93.89% of variants with allelic dosage $R^2$ ≥ 0.5. **e** Venn Diagram illustrating the loci identified by each reference panel. New loci are depicted in bold. As shown in this figure, only 67 of the 94 GWAS significant loci were identified by all four reference panels, while 27 of them (28.7%) were only identified by one, two, or three of the four panels.
We next tested all the 16 M variants for association with the 22 conditions available in the GERA cohort and five different inheritance models (Supplementary Figs. 4–25). This analysis identified 94 independent loci associated with 17 phenotypes at a usual genome-wide significance level \((p < 5.0 \times 10^{-8})\) of which 63 for 14 phenotypes were also experiment-wide significant \((p < 2.0 \times 10^{-8})\) after considering correction for the different models of inheritance (see “Methods”) (Supplementary Data 4). According to the GWAS Catalog, 68 of the 94 genome-wide significant loci had been previously reported to be associated with the same disease (Supplementary Data 5), whereas 26 of them correspond to previously unreported loci with associations across 16 phenotypes (Table 1).

Of these 26 new loci, 16 correspond to common, 3 to low-frequency, and 7 to rare variants. Only a fraction of the 26 new loci would have been genome-wide significant by using individual imputation panels (Fig. 1e), namely 19/26 using HRC, 14/26 using 1000G Phase 3, 13/26 using UK10K or 14/26 using GoNL. In addition, the lead marker for three of the novel signals was an indel, not covered by HRC, further confirming the benefits of combining multiple panels with our approach.

**Identification of recessive variants with large effects.** The implementation of refined GWAS strategies not only increases the number of associated variants, but also allows the identification of loci with large impact on the disease. Among the variants that were not detected under the additive model, and hence are expected to be missed by the majority of current GWAS, we highlight three variants with large recessive effects. First, an intronic indel in the CACNB4 gene, rs201654520, associated with a nearly 20-fold increase in risk for cardiovascular disease \((OR_{recessive} [95\% CI] = 1.75 \pm 0.0016)\). CACNB4 encodes the β4 subunit of the voltage-dependent calcium channel. This subunit contributes to the flux of calcium ions into the cell by increasing peak calcium current and triggering muscle contraction. Interestingly, an intronic single nucleotide polymorphism (SNP) within CACNB4, rs150793926, was associated with idiopathic dilated cardiomyopathy in African Americans\(^{26}\), but this variant is not in linkage disequilibrium (LD) with rs201654520 \((LD^2 = 0.0016\) for European ancestry and \(LD^2 = 0.0\) for African ancestry).

A second recessive variant with large effect, rs77704739, near the PELO gene, is associated with a fourfold risk for type 2 diabetes \((MAF = 0.036, OR [CI 95\%] = 4.3 [2.7–6.9], p = 1.75 \times 10^{-8})\). An independent signal that is about 112 K base pairs away (rs870992, LD \(r^2 = 0.0009\)) was previously associated with type 2 diabetes in the Greenlandic population, also with a recessive effect\(^{12}\). To provide insights into the underlying molecular mechanisms in disease, we interrogated comprehensive catalogs of genetic effects on gene expression: eQTLGen Consortium\(^{28}\) and GTEx\(^{29}\). The rs77704739 variant was significantly associated with gene expression of PELO in multiple tissues, including diabetes-relevant tissues such as adipose tissue, skeletal muscle, and pancreas. Colocalization analyses showed a probability higher than 0.8 in several tissues, including subcutaneous adipose tissue and skeletal muscle, suggesting this gene as the effector transcript (Fig. 2a, b, and Supplementary Data 6). In addition, we found that the lead variants in the PELO locus overlap with active promoter annotations in human pancreatic islets and open chromatin sites highly bounded by islet-specific transcription factors\(^{30,31}\) (Fig. 2c).

Third, a rare indel, rs557998486, located near the THUMP2 gene, is associated with age-related macular degeneration \((MAF = 0.009, OR = 10.5, p = 2.75 \times 10^{-8})\). Interestingly, the fact that we found no SNPs in LD with this lead indel further confirms the benefits of multiple reference panel imputation strategies that include alternative forms of variation. The lead indel rs557998486 overlaps DNase1 hypersensitivity sites in retinal and iris cell lines\(^{32}\), highlighting a candidate open chromatin region that is also predicted to be an enhancer assigned to the THUMP2 gene according to GeneHancer\(^{13}\). One of the variants with the highest LD with rs557998486 \((r^2 = 0.32)\) is associated with reduced expression of its nearest gene, THUMP2 \((\text{z-score} = -4.85, p = 1.25 \times 10^{-6})\), according to eQTLGen Consortium data.

Our empirical evaluation also demonstrates that our imputation approach is accurate for the new variants, including those with a large recessive effect, and that the combination of multiple reference panels increases the certainty of the imputed genotypes (Table 1 and Supplementary Data 7).

**Replication using UK Biobank and FinnGen.** We sought replication of previously unreported loci using UK Biobank, a prospective cohort of ~500 K individuals aged between 40 and 69\(^{3}3\). Given the high heterogeneity in phenotype definitions in UK Biobank compared to GERA, we tested for replication with the same phenotype and related traits (Supplementary Data 8). Compared to GERA, some of the conditions may not be ascertainment biased or have an age at onset later than the average age at ascertainment in UK Biobank \((56.52 \text{ years}^{34})\) which could affect the replication success. Despite these limitations, we tested the novel variants using the corresponding inheritance model and replicated 4 new loci with the same phenotype (Table 2).

The variant rs77704739 variant near PELO was associated with type 2 diabetes \((OR_{recessive} [95\% CI] = 1.9 [1.4–2.6], p = 4.95 \times 10^{-4})\) and metformin use \((OR_{recessive} [95\% CI] = 2.3 [1.6–3.4], p = 3.8 \times 10^{-5})\) in the UK Biobank\(^{5}\) (Supplementary Data 8), also only under the recessive model.

We further sought replication of the association within the CACNB4 gene with cardiovascular disease in FinnGen, a cohort of ~218 K Finnish individuals with an average age of 63, as it includes individuals with a higher average age \((63 \text{ vs} 56 \text{ in UK Biobank})\) and the risk of developing a cardiovascular disease is well-known to increase with age\(^{35}\). In addition, FinnGen has a precise and richer classification of this particular phenotype than UK Biobank. In brief, we tested rs201654520 for association with 47 cardiovascular endpoints. Of all the conditions tested, four (hypertensive heart disease, hypertensive heart and/or renal disease, heart failure, and right bundle-branch block) were nominally associated \((p < 0.05)\). All the associations had a direction of effect consistent with the effect observed in the GERA cohort (Supplementary Fig. 26a). Despite the high heterogeneity in the phenotype definitions between cohorts, we meta-analyzed the results from these endpoints from FinnGen with the result from cardiovascular disease phenotype from GERA, but none of them reach the genome-wide significance (see “Methods”) (Supplementary Fig. 26). We did not include UK Biobank in this meta-analysis as the equivalent phenotypes were not available or had less than 350 cases in UK Biobank, therefore, underpowered for a recessive analysis. Notably, when analyzing the association of rs201654520 with related quantitative traits we found that those who were homozygous for the high-risk allele had lower systolic blood pressure \((p = 4.1 \times 10^{-3}, \beta = -0.23)\) (Supplementary Data 8). While lower systolic blood pressure has been associated with increased risk of myocardial infarction in particular circumstances, this is not the typical direction of association, and therefore merits additional study\(^{36}\).

We also sought replication of the recessive association of rs557998486 near THUMP2 gene with macular degeneration in FinnGen. rs557998486 was associated with increased risk of macular degeneration in UK Biobank under the recessive model.
| Phenotype (gene/control) | CHR | SNP | Position | Minor allele | Dominant | Additive | Genotypic | Minor allele frequency, OR (CI 95%) | Minor allele frequency, OR (CI 95%) |
|--------------------------|-----|-----|----------|-------------|----------|----------|-----------|-------------------------------------|-------------------------------------|
| Major depression disorder | 5 | rs62250779 | G/A | 0.073 | 1.14 | 0.98 | 1.14 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| LMo7 | rs38546839 | C/G | 0.046 | 1.15 | 0.98 | 1.15 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| HIC1 | rs112963849 | C/A | 0.082 | 1.15 | 0.98 | 1.15 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| SNAR A12 | rs10112506 | A/G | 0.390 | 1.94 | 0.96 | 1.94 | 1.00 (1.00–2.00) | 1.00 (1.00–2.00) |
| Cancer (17,131/39,506) | 12 | rs123456789 | C/T | 0.005 | 1.14 | 0.98 | 1.14 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| DSC3 | rs109845678 | A/G | 0.390 | 1.94 | 0.96 | 1.94 | 1.00 (1.00–2.00) | 1.00 (1.00–2.00) |
| DCLRE1B | rs112963849 | C/A | 0.082 | 1.15 | 0.98 | 1.15 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| Cancer (17,131/39,506) | 20 | rs123456789 | C/T | 0.005 | 1.14 | 0.98 | 1.14 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| DSC3 | rs109845678 | A/G | 0.390 | 1.94 | 0.96 | 1.94 | 1.00 (1.00–2.00) | 1.00 (1.00–2.00) |
| DCLRE1B | rs112963849 | C/A | 0.082 | 1.15 | 0.98 | 1.15 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| Cancer (17,131/39,506) | 12 | rs123456789 | C/T | 0.005 | 1.14 | 0.98 | 1.14 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| DSC3 | rs109845678 | A/G | 0.390 | 1.94 | 0.96 | 1.94 | 1.00 (1.00–2.00) | 1.00 (1.00–2.00) |
| DCLRE1B | rs112963849 | C/A | 0.082 | 1.15 | 0.98 | 1.15 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| Cancer (17,131/39,506) | 12 | rs123456789 | C/T | 0.005 | 1.14 | 0.98 | 1.14 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
| DSC3 | rs109845678 | A/G | 0.390 | 1.94 | 0.96 | 1.94 | 1.00 (1.00–2.00) | 1.00 (1.00–2.00) |
| DCLRE1B | rs112963849 | C/A | 0.082 | 1.15 | 0.98 | 1.15 | 1.01 (0.83–1.25) | 1.00 (0.88–1.17) |
(OR [CI 95%] = 7.6 [1.5–37.3], \( p = 4.1 \times 10^{-2} \)), eye surgery (beta [CI 95%] = 1.6 [0.6–2.6], \( p = 1.17 \times 10^{-3} \)) (Supplementary Data 8), and with increased C-reactive protein, a known biomarker for macular degeneration\(^{37,38}\) (beta [CI 95%] = 1.1 [0.7–1.5], \( p = 1.15 \times 10^{-4} \)) (Supplementary Data 9). In FinnGen this variant was not significantly associated although it showed the same direction of effect. However, the meta-analysis did not reach the genome-wide significance (rs557998486 \( p = 9.6 \times 10^{-6} \))
and had a high heterogeneity (heterogeneity $I^2 = 87.1$, heterogeneity $p = 4.3 \times 10^{-4}$).

**Detection ranges of the different inheritance models.** Our findings provide an empirical overview of the detection range of five different inheritance models and show how each of them captures a fraction of the genetic variants associated with complex traits. Compared to current GWAS that usually only consider additive allelic effects, we found three different scenarios. Among all the 94 associated loci identified, 12 showed genome-wide significance only under the additive model, 62 under both additive and non-additive models, and 20 showed genome-wide significance only when non-additive tests were applied (Fig. 3a).

To further classify these variants, we tested whether any of the 62 variants associated with both additive and non-additive models deviate from additivity by a dominance deviation test\(^{11}\). Eleven of these 62 variants (17.7%) showed significant deviation from additivity (dominance deviation test $p < 0.05$). However, variants not showing a significant deviation from additivity may become significant for other models with larger sample sizes. Altogether, the dominance deviation test over the 93 autosomal loci identified 62 (66%) additive and 24 (25.5%) non-additive associations, and 8 undetermined. Based on the smallest GWAS $p$-value, we further classified non-additive associations into 9 recessive, 13 dominant, 8 heterodominant, and 7 genotypic (Supplementary Data 4).

We also observed that each of the available models for association testing has a different range of detection. To identify the 94 genome-wide associated loci, the additive test, as expected, was the most sensitive model (74 loci), followed by the genotypic (59 loci), the dominant (56 loci), the recessive (43 loci), and the heterodominant (32 loci). When considering known loci, 48 of the 94 were identified as non-additive as associations by more than one model in our analysis, and almost half of these (22 loci) with all three credible set variants are located in open chromatin sites in human pancreatic islets, one of them classified as an active promoter and one highly bounded by pancreatic islet-specific transcription factors, such as PDX1, NKX2.2, NKX6.1, and FOXA2.

The **GUIDANCE framework.** The complete methodology described here and used for the analysis of the GERA cohort was integrated into a framework, called GUIDANCE. GUIDANCE allows the analysis of genome-wide genotyped data in a single execution in distributed computing infrastructures without the need for extensive computational expertise or constant user intervention. The GUIDANCE workflow requires quality-controlled genotyped data as input and provides association results, graphical outputs, and statistical summaries. Integrating state-of-the-art tools with in-house code written in Java, bash, and R\(^{39}\), GUIDANCE efficiently performs large-scale GWAS, including (1) the pre-phasing of haplotypes, (2) the imputation of genotypes using multiple reference panels, (3) the association testing for different inheritance models and the integration of the results from different panels, (4) a cross-phenotype analysis when more than one phenotype is available in the cohort (Supplementary Data 10), and finally, (5) the generation of summary statistics tables and graphic representations of the results (Supplementary Fig. 27), for both the autosomes and the X chromosome. While GUIDANCE can be executed as a standalone compact program it can also be used in modules (Supplementary Fig. 28), which makes it adaptable to existing frameworks and provides a higher level of control to users.

GUIDANCE runs in distributed computing platforms, including the cloud, without requiring a broad background in these environments. This is feasible since GUIDANCE was implemented on top of the COMP Superscalar Programming Framework (COMP3\(^{40}\) ). The GUIDANCE workflow was implemented as a sequential Java program containing the calls to the GWAS tools, encapsulated in Java methods, and selected as tasks, while COMP3 controls the execution of those tasks on the underlying distributed infrastructure. The source code, the pre-compiled binaries, and the documentation to use GUIDANCE are available at http://cg.bsc.es/guidance.

**Discussion**

Current genetic association studies are undergoing fundamental strategic and methodological changes to identify new associations. The gradual incorporation of WES and WGS, together with the continuous efforts to increase the sample size, impose computational and methodological challenges that cannot always be overcome. In addition, the inclusion of the X chromosome, and non-additive models during association testing, despite increasing the computational burden, can also lead to the identification of new associations. In this study, we demonstrate the value of applying a comprehensive GWAS strategy, including denser imputation strategies, the X chromosome, and non-additive association tests, to an existing large-scale genetic resource, the GERA cohort. We show that by applying more innovative imputation protocols we increased the number and the type of variants tested for association, including low-frequency and rare SNPs as well as alternative forms of variation, such as indels. Our analysis of the GERA cohort shows that between 13 and 20 of the genome-wide significant associations (14–21%) would not have
| CHR | rsID (alleles) (MAF) | Best model | Stage 1. GERA Discovery | Stage 2. UK Biobank replication | Stage 1 - Stage 2 Meta-analysis |
|-----|---------------------|------------|-------------------------|--------------------------------|-------------------------------|
|     |                     | Phenotype (cases/controls) | Additive | Best model | Additive | Lowest p-value | Additive | Lowest p-value | Additive | Lowest p-value |
| 18  | rs2014497 (A/G) (0.008) | Additive Cancer (17,131/39,506) | OR (CI 95%) P-value OR (CI 95%) P-value OR (CI 95%) P-value | Self-reported: chronic lymphocytic (257/360,904) OR (CI 95%) P-value Self-reported: kidney/renal cell cancer (473/360,668) OR (CI 95%) P-value C69 Malignant neoplasm of eye and adnexa (1/46/30,048) OR (CI 95%) P-value | 2.13 (1.14–3.97) 10^-2 2.13 (1.14–3.97) 10^-2 3.50 (1.33–9.74) 10^-5 1.75 (1.07–2.86) 10^-5 1.71 (1.32–2.13) 10^-5 |
| 1   | rs2494196 (C/A) (0.274) | Additive Hernia abdominopelvic (6,291/50,346) | OR (CI 95%) P-value OR (CI 95%) P-value OR (CI 95%) P-value | Self-reported: umbilical hernia (328/360,813) OR (CI 95%) P-value K40 Inguinal hernia (13,365/347,829) OR (CI 95%) P-value K41 Femoral hernia (475/360,719) OR (CI 95%) P-value K42 Umbilical hernia (2623/358,571) OR (CI 95%) P-value K43 Ventral hernia (2470/358,224) OR (CI 95%) P-value | 2.03 (1.08–1.18) 10^-8 2.31 (1.21–4.41) 10^-10 1.42 (1.21–1.67) 10^-5 1.09 (1.06–1.12) 10^-10 1.09 (1.06–1.12) 10^-10 1.24 (1.26–1.64) 10^-7 1.07 (1.01–1.14) 10^-7 1.24 (1.26–1.64) 10^-7 |
| 2   | rs557998486 (T/G) (0.009) | Additive degeneration (3685/52,952) | OR (CI 95%) P-value OR (CI 95%) P-value OR (CI 95%) P-value | Eye problems/disorders: macular degeneration (2726/155,164) OR (CI 95%) P-value | 2.75 (1.27–5.58) 10^-8 1.07 (0.81–1.41) 10^-1 1.05 (0.81–1.41) 10^-1 1.26 (0.98–1.64) 10^-1 |
| 5   | rs77704739 (T/C) (0.036) | Additive Type 2 diabetes (6,967/49,670) | OR (CI 95%) P-value OR (CI 95%) P-value OR (CI 95%) P-value | Self-reported: diabetes (14,114/347,027) OR (CI 95%) P-value | 1.75 (1.05–2.63) 10^-8 4.32 (1.70–6.92) 10^-3 4.32 (1.70–6.92) 10^-3 1.88 (1.85–2.90) 10^-4 4.95 (2.93–8.26) 10^-4 |

Chr chromosome, Position position hg19, Allele non-effect allele/effect allele, MAF minor allele frequency, OR odds ratio.

*Odds ratio calculated using the recessive allele frequency-based test (RAFT).

*Obtained through a meta-analysis with UK Biobank using the expected method from SMPTEST.

*Obtained using METAL method SAMPLEIZE to combine the p-values taking into account the sample size and direction of effect.
been identified when using a single reference panel. Likewise, our analysis in the GERA cohort demonstrates that 21% of the associations would have been missed by only testing the additive model. Overall, 27.6% of associations would not have been identified by applying the commonly used HRC and additive model association testing. The inclusion of the X chromosome within the study allowed us to identify an intronic variant (rs67646851) associated with dyslipidemia in the RGAGI gene among the previously known associated loci since a variant in LD with our top variant (rs5985471, LD $r^2$ = 0.92 for European ancestry) has
already been associated with low-density lipoprotein cholesterol\textsuperscript{20}. This supports the value of analyzing the X chromosome, which is not considered in the majority of GWAS.

In this study, we show the potential of identifying large effect recessive associations by maximizing the use of current reference panels and testing different inheritance models, as exemplified by the associations with type 2 diabetes, cardiovascular disease, and macular degeneration with variants near \textit{PELO}, \textit{CACNB4}, and \textit{THUMPD2}, respectively. This strategy opens new avenues for future analyses in large-scale biobanks, as demonstrated with our power calculations, which show that even the largest available GWAS meta-analyses or biobanks would not have enough power to identify these associations using only the additive model. For example, the \textit{CACNB4} gene, associated with cardiovascular disease, would require a sample size equivalent to 370,000 individuals when using the additive test, 17 times larger than the required sample size under a recessive analysis. After considering all the supporting evidence illustrated with many examples in this study, the results suggest that these new associations deserve future validations and follow-up analysis. Therefore, this study demonstrates the importance of a comprehensive analysis including non-additive models when performing GWAS, which can increase the discovery not only on GWAS relying on genotyping array data, but also on WES or WGS association studies.

The inclusion of non-additive associations can also have an impact on the construction of polygenic risk scores. Current polygenic risk scores (PRS) are calculated summing risk alleles weighted by effect sizes from GWAS results, which have typically tested only the additive model in the association test. Hence, large-scale genome-wide association data accounting for different models of inheritance and including both SNPs and alternative forms of variation, such as indels, will also be essential to develop more accurate genome-wide PRS, which would weight each of the genotype carriers appropriately, rather than weighting the heterozygous halfway between the homozygous of the effect and alternate alleles.

To easily apply this strategy to genetic studies we present GUIDANCE, a standalone and easy-to-use application that allows an efficient and comprehensive GWAS analysis in different computing platforms, such as cloud and high-performance computing architectures. GUIDANCE is designed to allocate an unlimited number of reference panels. This can be useful for GWAS performed in specific diverse populations where the addition of a population-specific reference panel alongside the commonly used ones could be an advantage. This feature can also be useful when incorporating reference panels that have a better ascertainment of other type of variation, such as structural variants, with the most commonly used ones. In addition, it is possible to launch the execution by steps, incorporating the previously obtained results, avoiding repeating all the computations. In a moment where the community is facing computational and methodological challenges due to the growing complexity and size of genetic datasets, the availability of robust and complete analysis platforms can improve the efficiency of genetic studies, standardizing analysis strategies among meta-analysis cohorts to ensure consistency.

Finally, to share our results with the community and to promote the analysis of non-additive inheritance models in GWAS, a public searchable database including additive and non-additive summary statistics for 16 M of variants and 22 phenotypes is available at the Common Metabolic Diseases Knowledge Portal (https://cmdkp.org/), and full summary statistics both at the Common Metabolic Diseases Knowledge Portal and at http://cg.bsc.es/guidance.

**Methods**

**Empirical evaluation of the imputation strategy.** To empirically evaluate our imputation strategy, we extracted the corresponding genotypes in UK10K for each variant present in the genotyping array used in GERA to build an in silico array of 599,208 genotyped variants for the 3781 UK10K individuals. We then imputed the data using the 399,208 variants as a backbone, using 1000G, GoNL, and HRC as reference panels. We used IMPUTE2 to impute the genotypes. We also combined the imputed results based on the highest imputation accuracy according to the IMPUTE2 info values as we did with the GERA cohort. We transformed the IMPUTE2 genotype probabilities into genotype dosages and calculated the r-squared correlation ($R^2$) between the imputed genotype dosages and the sequenced genotype dosages from UK10K, termed allelic dosage $R^2$ (Supplementary Fig. 1).

**GUIDANCE workflow description.** By combining and integrating state-of-the-art GWAS analysis tools into the COMP Superscalar programming Framework (COMPSs), we developed GUIDANCE, a standalone application that performs haplotype phasing, genome-wide imputation, association testing, and PheWAS analysis of large GWAS datasets (Supplementary Fig. 27).

As shown in Supplementary Fig. 27, GUIDANCE’s workflow starts with genotype data and ends with providing association results, graphical outputs, and a statistical summary. Once everything is settled in the GUIDANCE configuration file, GUIDANCE performs an efficient two-stage imputation procedure, by pre-phasing the genotypes into whole haplotypes followed by genotype imputation itself\textsuperscript{81}, SHAPEIT\textsuperscript{20} or EAGLE\textsuperscript{22} and IMPUTE2\textsuperscript{22} or MINIMAC\textsuperscript{44} can be used for pre-phasing and genotype imputation, respectively. In addition, GUIDANCE accepts one or multiple reference panels, allowing the integration of the results obtained from all panels by selecting for each variant the genotypes from the reference panel that provides the highest imputation accuracy according to the IMPUTE2 info score or MINIMAC2 $r^2$ (Supplementary Fig. 29). GUIDANCE also performs a post-imputation quality control to eliminate low-quality imputed variants under the basis of the IMPUTE2 info score or MINIMAC2 $r^2$ and the MAF.

After genotype imputation and post-imputation quality control, GUIDANCE applies SNPISTEST for association testing, where additive, dominant, recessive, heterodominant, and genotype models can be analyzed. Here, the user can decide to include several covariates for the association test, such as principal components to adjust for population stratification, or any other confounders. GUIDANCE also allows testing for multiple phenotypes or for a single phenotype with different covariates in the same execution. After association testing, variants are filtered by the deviation from Hardy–Weinberg equilibrium (HWE) $p$-value. Finally, GUIDANCE generates summary reports for each trait with all the inheritance models tested in the association and the corresponding graphical representation, i.e., Manhattan and quantile–quantile (Q–Q) plots (Supplementary Fig. 4–25), also providing a matrix identifying cross-phenotype associations (Supplementary Data 2).

GUIDANCE can be executed as a standalone compact program or as independent modules (see Supplementary Fig. 28 for a list of independent modules) to facilitate the use of GUIDANCE into existing frameworks.
Further details can be found in the configuration file of the GUIDANCE execution at http://cg.bsc.es/guidance. Specific documentation to use this framework is available at http://cg.bsc.es/guidance, as well as the source code and the pre-compiled binaries that are available in the Download section.

The analysis of GERA cohort

GERA cohort description. GERA cohort data was obtained through dbGaP under accession phs000674.v1.p1. For further information about the specific phenotypes (ICD-9-CM codes) included in GERA, visit its website on dbGaP (https://www.ncbi.nlm.nih.gov/projects/gp/cgi-bin/GetaP?db=phs200380). The Resource for Genetic Epidemiology Research on Aging (GERA) Cohort was created by a RC2 Grand Opportunity grant that was awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics (AG036607; Schafer/Ritch, PIs). The RC2 project enabled genotyping SNPs across the genome (GWAS) to be performed on 200 K adults who were members of the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC), and participating in its RPGEH. The resulting GERA cohort is composed of 42% of males, 58% of females, and ranges in age from 18 to over 100 years old with an average age of 63 years at the time of the RPGEH survey (2007). The individuals are non-European, while 81% are described as white non-Hispanic participants. After an explicit requirement of consent by email, data from 78,486 participants were deposited in dbGaP, with similar demographic characteristics to those of the initial genotyped cohort.

Quality control. A subset of 62,281 subjects of European ancestry underwent quality control analyses. A three-step quality control protocol was applied using PLINK45,46, and included two stages of SNP removal and an intermediate stage of sample exclusion.

The exclusion criteria for genetic markers consisted of proportion of missingness ≥ 0.05, HWE p ≤ 1 × 10⁻⁶ for all the cohort, and MAF ≤ 0.001. This protocol for genetic markers was performed twice, before and after sample exclusion.

For the individuals, we considered the following exclusion criteria: gender discordance, subject relatedness (pairs with PI-HAT ≥ 0.125 from which we removed the individual with the highest proportion of missingness), sample call rates ≥ 0.02, and population structure showing more than four standard deviations within the distribution of the study population according to the first seven principal components (Supplementary Fig. 30). After QC, 56,637 subjects remained for the analysis (Supplementary Data 1).

Analyzing GERA cohort using GUIDANCE. GUIDANCE pre-phased the genotypes to whole haplotypes with SHAPEIT2, and then performed genotype imputation with IMPUTE2 using 1000G phase 3, UK10K, GoNL, and HRC as reference panels. After filtering variants with an info score < 0.7 and a MAF < 0.001, we tested for association with additive, dominant, recessive, heterodominant, and genotypic logistic regression using SNPTEST, and including seven derived principal components and age as covariates. To maximize power and accuracy, we combined the association results from the four reference panels by choosing for each variant, the genotypes from the reference panel that provided the best IMPUTE2-info score.

As chromosome X we restricted the analysis to non-pseudoautosomal (non-PAR) regions. For the haplotype phasing of chromosome X, we used the --chx flag required for SHAPEIT to only phase female samples since males only have one X chromosome, and to impute missing data in male samples. We did the same imputing genotypes using IMPUTE2 as it requires the --chx flag alongside the sample file with the sex information. In the association test, we used method newml from SNPTEST to ignore samples with missing sex or males encoded wrongly (males should be coded 0/1, as homozygote females), and to assume a complete inactivation of one allele in females and equal effect size between males and females. For heterogeneity between males and females, and to allow a complete inactivation of the X chromosome in females, we used --stratify on sex to separate the effects and the baselines of males and females, accounting for hemizygosity for males, while for females, we followed an autosomal model.

Finally, we excluded variants with HWE controls p < 1 × 10⁻⁶ and with a case count for homozygous for the alternative allele below three in the final results for the recessive and genotypic model, as we observed a trend for genomic inflation and deflation in the recessive and genotypic model before removing these variants.

Identification of known and new associated loci. After the association test, GUIDANCE provided a list of variants that passed the p-value threshold specified in the configuration file (i.e., p ≤ 5.0 × 10⁻⁸). Using the IRanges R package57, all the genome-wide significant variants were collapsed into ranges (500 kb) that define each associated locus.

To distinguish between known or new associated regions, for each top variant we looked for any proxy variant with an LD r² > 0.35 in the GWAS catalog (accession 5 September 2019) associated with the same phenotype or a related one (for example, bone mineral density, blood pressure, or diabetes). For each blood pressure phenotypes for osteoporosis, dyslipidemia or hypertension, respectively). HLA regions at chromosome X were excluded since the particularities of these regions required further detailed studies on their LD pattern. Proxies were selected using LDlink (https://ldlink.nci.nih.gov/).58

We defined an empirically significant p-value cutoff of p < 2.0 × 10⁻⁷ by applying the Bonferroni correction for 2.5 effective test (5.0 × 10⁻⁷-2.5×10⁻⁷ effective test). This factor of 2.5 was obtained from a simulation study when four genetic models (additive, dominant, recessive, and genotypic) are used59 since the genetic models are not independent. However, a new simulation study including the heterodominant model should be done for a more accurate effective number of tests.

Replication with UK biobank

Phenotype curation. UK Biobank participants agreed to provide detailed information about their lifestyle, environment, and medical history, to donate biological samples (for genotyping and for biochemical assays), to undergo measures, and to have their health followed (http://www.ukbiobank.ac.uk/).

When collecting and analyzing a wide range of phenotypes from the UK Biobank, a central challenge was the curation and harmonization of the vast array of categorizations, variable scales, and follow-up responses. Fortunately, to this end, the Phenome Scan Analysis Tool (or PHESANT: https://github.com/MRCIEU/PHESANT) performs much of the transformations and recordings required to generate meaningful, interpretable phenotypes.

We have made further adjustments based on user feedback, owing to the value of transparent information in generating our phenotype guidelines. Applying these changes to the PHESANT source code, phenotypes were parsed using our modified version (github.com/asth/phenosets/PHESANT) on a virtual machine on the Google Cloud Platform.

First, we restricted to the subset of European individuals, before passing the resultant phenotypic data to PHESANT. The ‘variable list’ file and ‘data-coding’ file, whose formats are defined in the original version of PHESANT were updated as new phenotypes were added in the latest UK Biobank release. Recordings of variables, and inherent orderings of categorical variables, are defined in the ‘data-coding’ file. The ‘Excluded’ column of the ‘variable list’ file defines the collection of variables that we do not wish to interrogate.

A high-level overview of the PHESANT pipeline, our default, and the associated short flags for the phenoset5 syntax code are displayed in Supplementary Fig. 31. In addition to the inverse-rank normalization applied to the collection of continuous phenotypes, we also consider the raw version of the continuous phenotype, with no transformation applied to the data.

Curation of the ICD10 codes was carried out separately for computational effects. For the ICD10 phenotype, individuals are assigned a vector of ICD10 diagnoses. We truncated these codes to two digits, and assigned each individual to either case or control status for that ICD10 code in turn by checking if their vector contains that code. Throughout, we assumed the data contained no missingness, so the sum of cases and controls throughout was the number of individuals in our ‘European’ subset of the UK Biobank data. As in the PHESANT categorical (multiple) phenotypes, ICD10 code case/control phenotypes were removed if ≤50 individuals had the diagnosis.

Association testing and meta-analysis for UK Biobank phenotypes. We performed the association testing for the curated phenotypes as implemented in SNPTEST for additive, dominant, recessive, heterodominant, and genotypic inheritance models, as it has been described in the “Analyzing GERA cohort using GUIDANCE” section. For all genotypic variants identified in the discovery stage, we assigned the recessive model after we identified it as the underlying model.

After the association testing, we filtered and ordered all the phenotypes based on the p-value for the best model of inheritance obtained from the GERA cohort analysis, with special consideration to equivalent phenotypes or related traits.

With the association testing results of both GERA cohort and UK Biobank, we meta-analyzed the results using METAL53. We use the inverse variance-weighted fixed-effect model for all the variants except for the rs557998486 variant associated with macular degeneration, since its beta, calculated with the em method from SNPTEST, was inflated. Therefore, we performed a sample size-based meta-analysis, which converts the direction of the effect and the p-value into a z-score. For biomarkers, only the results from the first visit were taken into account since <10% of the cases were present in the second visit.

Association testing and meta-analysis with FinnGen. We used SAIGE52 for recessive association testing using sex, age, PC1–10, and batch as covariates. We analyzed FinnGen release 5 that contains 218,792 individuals with a median age 62.6 and a mean age 59.8.

For the cardiovascular disease endpoints, we meta-analyzed the results using meta R package53. For macular degeneration, we meta-analyzed the results using METAL as described in the previous section.

Dominance deviation test. To detect genuine differences between additive and non-additive signals, we performed a dominance deviation test for all 93 autosomal genome-wide significant loci.

Dominance deviation test was performed using a logistic regression analysis using PLINK (v1.90b6, www.cog-genomics.org/plink1/1.9). Sex, age, and the first 7 PCs were included as covariates.
Definition of 99% credible set of PELO locus. For the PELO locus, the fraction of aggregated variants that have a 99% probability of containing the causal one was identified. The credible set of variants for the region was defined with a Bayesian refinement approach, considering variants with an $r^2 > 0.1$ with the leading one. For each variant within the PELO locus, the credible set provides a posterior probability of being the causal one. The approximate Bayes factor (ABF) for each variant was estimated as:

$$ABF = \frac{\sqrt{1 - r^2 e^{r^2/2}}}{0.04},$$

where

$$r = \frac{z}{\sqrt{SE^2 + 0.04}}.$$  

The $\beta$ and the SE result from a logistic regression model testing for association. The posterior probability for each variant was calculated as:

$$\text{Posterior probability} = \frac{ABF}{P},$$

where ABF corresponds to the approximate Bayes’ factor for the marker $i$, and $P$ represents the sum of all the ABF values enclosed in the interval. As commonly employed by SNPTEST, this calculation assumes that the prior of the $\beta$ is a Gaussian with mean 0 and variance 0.04.

Finally, the cumulative posterior probability was calculated after ranking the variants according to the ABF in decreasing order. Variants were included in the 99% credible set of the region until the cumulative posterior probability of association got over 0.99.

Gene expression and functional characterization. The eQTLGen Consortium (https://www.eqtlgen.org/cis-eqtls.html, last access on July 2019) and GTEx portal (http://cg.bsc.es/gera_summary_stats) can be cited with the DOI 10.5281/zenodo.4446121. It contains example configuration files, explanations on how to execute the framework, a containerized version of the pipeline, and the source code. More detailed instructions can be found on its website (http://cg.bsc.es/gera_summary_stats).

Received: 11 May 2020; Accepted: 9 February 2021; Published online: 23 April 2021

References

1. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42, D1001–D1006 (2014).

2. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).

3. Bonas-Guarch, S. et al. Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. Nat. Commun. 9, 321 (2018).

4. Talou, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021).

5. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–208 (2019).

6. Tam, V. et al. Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. 20, 467–484 (2019).

7. Ziaei, Z. et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. Am. J. Hum. Genet. 96, 377–385 (2015).

8. Salanti, G. et al. Underlying genetic models of inheritance in established type 2 diabetes associations. Am. J. Epidemiol. 170, 537–545 (2009).

9. Lettre, G., Lange, C. & Hirschhorn, J. N. Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet. Epidemiol. 31, 356–362 (2007).

10. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. Nat. Rev. Genet. 7, 277–282 (2006).

11. Wood, A. R. et al. Variants in the FTO and CDAK1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively. Diabetologia 59, 1214–1221 (2016).

12. Grarup, N. et al. Identification of novel high-impact recessively inherited type 2 diabetes risk variants in the Greenlandic population. Diabetologia 61, 2005–2018 (2018).

13. Moltke, I. et al. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature 512, 190–194 (2013).

14. Steinhardt, C. T., Stocks, A., O’Donnell, C. J., Libiger, O. & Schuler-G�iger, N. A variant in CDAK1 influences insulin response and risk of type 2 diabetes. Nat. Genet. 39, 770–775 (2007).

15. Lenz, T. L. et al. Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat. Genet. 47, 1085–1090 (2015).

16. Goyette, P. et al. High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. Nat. Genet. 47, 172–179 (2015).

17. Hoffmann, T. J. et al. Imputation of the rare HOBX13 G48E mutation and cancer risk in a large population-based cohort. PLoS Genet. 11, e1004930 (2015).

18. Boomsma, D. I. et al. The Genome of the Netherlands: design, and project goals. Eur. J. Hum. Genet. 22, 221–227 (2014).

19. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46, 818–825 (2014).

20. UK10K Consortium. et al. The UK10k project identifies rare variants in health and disease. Nature 526, 82–90 (2015).

21. 1000 Genomes Project Consortium. et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).

22. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48, 1279–1283 (2016).

23. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).

24. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. 50, 1505–1513 (2018).

25. Fritsch, 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).

26. Xu, H. et al. A genome-wide association study of idiopathic dilated cardiomyopathy in African Americans. J. Pers. Med. 8, 11 (2018).

27. Hill, W. G. & Robertson, A. Linkage disequilibrium in finite populations. Theor. Popul. Biol. 33, 266–271 (1988).

28. Võsa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. Preprint at bioRxiv https://doi.org/10.1101/447367 (2018).

29. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585 (2013).

30. Pasqualli, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat. Genet. 46, 136–143 (2014).
Acknowledgements
This work has been sponsored by the grant SEV-2011-00067 and SEV2015-0493 of Severo Ochoa Program, awarded by the Spanish Government, by the grant TIN2015- 65316-P, awarded by the Spanish Ministry of Science and Innovation, and by the Generalitat de Catalunya (contract 2014-SGR-1051). This work was supported by an EFSID/Lilly research fellowship. Josep M. Mercader was supported by a Sara Borrell Fellowship from the Instituto Carlos III, Beatriz de Pinós fellowship from the Agency for the Promotion of Research and Innovation of Catalonia (AGAUR), and the Generalitat de Catalunya Innovative and Clinical Translational Award 1·19·ICTS·068. Silvia Bonas was supported by FI-DGR Fellowship from FIDGR 2013 from Agencia de Gestió d’Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya), and a ‘Juan de la Cierva’ postdoctoral fellowship (MINECO/FJCI-2017-32909). Cecilia Salvorre received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433. Cristian Ramon-Cortes pre-doctoral contract is financed by the Spanish Ministry of Science, Innovation, and Universities under contract BES-2016-07679. Elizabeth G. Atkinson was supported by the National Institutes of Mental Health (grants K01MH121659 and T32MH017119). Jose Florrez was supported by NIH-NIDDK award K24 DK115050. This study made use of data generated by the UK10K Consortium, derived from samples UK10K COHORT IMPUTATION (EGAS00001000713). A full list of the investigators who contributed to the generation of the data is available at www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. This study made use of data generated by the ‘Genome of the Netherlands’ project, which is funded by the Netherlands Organization for Scientific Research (grant no. 104001001). The data were made available as a Rainbow Project of BBMRI-NL. Samples were contributed by Lifelines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.healthy-ageing.nl; http://www.langleven.eu), the Netherlands Twin Registry (NTR; http://www.twelingenregister.org), the Rotterdam studies (http://www.eramus-epidemiology.nl/rotterdamstudy) and the Genetic Research in Isolated Populations program (http://www.epib.nl/research/geneticsci/genetics). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI). This study also made use of data generated by the Haplotype Reference Consortium (HRC) accessed through the European Genome-phenome Archive at the European Bioinformatics Institute with the accession numbers EGAD00001002729, after a form agreed by the Barcelona Supercomputing Center (BSC) with WTSL. This research has been conducted using also the UK Biobank Resource (application number 31063 and 27892). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 07/16/2019. We acknowledge PRACE for awarding us access to both MareNostrum supercomputer from the Barcelona Supercomputing Center, based in Spain at Barcelona, and the SuperMUC supercomputer of the Leibniz Supercomputing Center (LRZ), based in Garching at Germany (proposals numbers 2016143358 and 2016163985). The technical support group from the Barcelona Supercomputing Center is gratefully acknowledged. Finally, we thank all the Computational Genomics group at the BSC for their helpful discussions and valuable comments on the manuscript. We also acknowledge Elias Rodriguez Fox for designing the GUIDANCE logo.

Author contributions
M.G.-M., J.M.M., and D.T. conceived and planned the main analyses. M.G.-M., J.M.M., and D.T. wrote and edited the manuscript. S.B.-G. designed and performed the quality control. M.G.-M. performed the main analysis, and with collaboration of R.A., M.P., and J.C.F., contributed with UK10K HARD data analysis. S.B.-G. and M.K. contributed with FinnGen data and analysis. J.M.M. and D.T. supervised the study. All authors reviewed and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21952-4.

Correspondence and requests for materials should be addressed to J.M.M. or D.T.

Peer review information Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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