Rare-variant collapsing analyses for complex traits: guidelines and applications

*Gundula Povysil[1], Slavé Petrovski[2,3], Joseph Hostyk[1], Vimla Aggarwal[1], Andrew S. Allen[4] and David B. Goldstein[1]**

Abstract | The first phase of genome-wide association studies (GWAS) assessed the role of common variation in human disease. Advances optimizing and economizing high-throughput sequencing have enabled a second phase of association studies that assess the contribution of rare variation to complex disease in all protein-coding genes. Unlike the early microarray-based studies, sequencing-based studies catalogue the full range of genetic variation, including the evolutionarily youngest forms. Although the experience with common variants helped establish relevant standards for genome-wide studies, the analysis of rare variation introduces several challenges that require novel analysis approaches.

Penetrant alleles
Alleles highly associated with a trait; the more penetrant the allele, the higher the percentage of individuals with that allele who also express a disease phenotype.

Deleterious variation
Genetic variation that is predicted to disrupt gene function and therefore lead to reduced fitness.

[1]*Institute for Genomic Medicine, Columbia University Irving Medical Center, Columbia University, New York, NY, USA.*

[2]*Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK.*

[3]*Department of Medicine, The University of Melbourne, Austin Health and Royal Melbourne Hospital, Melbourne, Victoria, Australia.*

[4]*Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA.*

**e-mail: dg2875@cumc.columbia.edu*

https://doi.org/10.1038/s41576-019-0177-4

Before next-generation sequencing (NGS), complex trait genetics focused on common variant contributions to disease risk using a genome-wide association study (GWAS) design[1]. These studies successfully identified a large number of significantly associated loci across a number of complex diseases. However, as study sample sizes have grown, newly identified variants have had smaller effects on risk. In fact, it has recently been argued that if enough individuals are genotyped, the common variants that are found to be significantly associated with several complex traits would be spread broadly and very densely across the genome (omnigenic model)[2]. This model is contrary to the original hope of GWAS: that the variants would cluster among, and hence implicate, key biological pathways (in a classical polygenic model). The clear implication of the omnigenic model is that except for some big hits that affect core genes, many common variant association signals emerging from GWAS would not generally be expected to provide a mechanistic understanding of disease or guidance regarding optimal clinical management for individual patients[3]. By contrast, rare-variant studies usually detect variants with larger effect sizes that may implicate additional core genes and therefore might lead to more direct insights into disease biology. A key limitation of studies based on genotype data has been that markers had to be predesigned in the genotyping microarrays and were limited in number. Therefore, the focus was typically on more-common variants. Advances in NGS technology have transformed human and medical genetics in the past decade by enabling a more complete assessment of an individual's genetic variation, including the rare (younger) alleles[4,5].

The contribution of rare variants to a range of human phenotypes is already well established, with many disorders being explained by individual, highly penetrant alleles with reduced reproductive fitness[4,5]. This reduced reproductive fitness prevents the variants from becoming common in the general population, highlighting again why a focus on rare variants is important. The earliest applications of exome sequencing in human genetics focused on the diagnostic interpretation of individual patients with a presumed 'simple' genetic condition — an application that has been very successful[6–12]. One question that remained largely unexplored until more recently, however, was the role that rare variation plays among the more common and complex traits. The need to address this question saw the emergence of rare-variant analytical approaches aimed at capturing rare-variant information across a gene or other defined genomic units. These approaches allow the identification of genes containing an excess of rare and presumably deleterious variation among cases ascertained for complex disease traits, relative to controls.

As NGS technologies became more high-throughput and costs continued to drop, the application of whole-exome sequencing (WES) and whole-genome sequencing (WGS) studies became more attractive among common complex disorders. Recent rare-variant studies have not only led to the identification of genes that show definitive genome-wide significant association with disease, but also provided insight into key issues that have otherwise been difficult to address, including identifying specific variants that contribute to disease risk, evaluating the relative contributions of individual disease genes

to overall disease burden and assessing aspects of genetic architecture, including comparing the contributions of different allele frequencies and variant effect classes[13–15].

A common theme across recent NGS studies is that rare variants, and in particular those found to be 'ultra-rare' in the population — that is, unobserved in available reference cohorts and likely to be very young in origin — also play an important role in the genetic architecture of complex disorders. In this Review we focus on the gene-based collapsing approach, in which variants that satisfy specific criteria (qualifying variants) are binned together as equivalent, as a simple yet demonstrably effective approach to identifying rare-variant contributions to disease. We summarize important lessons learned from the application of collapsing analyses to a range of phenotypes, including how applications have been optimized to improve genetic risk signal detection. Additionally, we describe elaborations currently being developed, such as incorporating regional intolerance within sub-regions of genes as an additional source of information[16–20], aggregating information across multiple genes in order to understand the importance of various biological pathways, and applying genetic models beyond the single-gene-dominant models that have largely been considered to date. Finally, we discuss how analogous approaches can be applied to whole-genome sequence data, emphasizing the challenges related to defining the units within which to collapse variants and how to enrich for functional variants within those regions.

Introduction to rare-variant collapsing

Conceptually, gene-based rare-variant approaches work optimally when an expectation of allelic heterogeneity exists among the one or many disease-associated genes[21]. In these situations, each individual causal allele is expected to explain only a very small fraction of the cases under study, but different variants in the same gene may have a larger cumulative contribution. An intuitive example for this is provided by haploinsufficiency-mediated disorders. In haploinsufficient disease genes, the number of different alleles that confer equivalent risk is expected to be large and generally recognizable: any loss-of-function (LOF) allele, whether in an essential splice site, a frameshift or a stop mutation, will result in haploinsufficiency and, hence, disease. In such a case, it is reasonable to flag the presence (or absence) of any LOF variant and simply to test whether an increased number of cases have an LOF variant, relative to controls. In this context, the control sample is extremely important, as it reflects the empirical background variation rate for the suspicious class of variation in the test gene[21].

Qualifying variant. In most cases, the successful application of collapsing analyses depends on optimizing parameters in order to focus on the class of variation that will enrich for variants that confer risk and to reduce the impact of neutral background variation. A qualifying variant is one that is observed in the cases and controls being tested and passes a collection of filters[13]. Commonly applied filters include sequencing-based quality metrics, predicted variant effects, predictions of deleteriousness and — possibly the most important metric — population allelic frequencies. Specifically, for diseases under strong negative selection, restricting the analyses to the rarest variants has shown strong enrichments for causal variants[14].

Gene-based collapsing approach. The conventional gene-based collapsing approach uses the protein-coding boundaries of genes to evaluate whether there is a significant difference in the counts of cases versus controls who carry at least one qualifying variant[13]. This approach is particularly useful in traits for which a simple genetic model explains a proportion of the case population. Each gene is individually assessed for significant differences in counts of case and control individuals carrying a qualifying genotype. Given the approximately 19,000 protein-coding genes, the conventional exome-wide multiplicity-adjusted significance threshold is assigned as $\alpha = (0.05/19{,}000) \approx 2.6 \times 10^{-6}$. As the number of hypotheses is clearly defined by the number of genes, exome-wide significance accounting for all independent hypotheses is the appropriate approach, as opposed to replication[22]. If multiple collapsing models are applied, as is usually the case, the significance threshold needs to be further divided by the number of models tested.

Other rare-variant association methods. Many approaches have been suggested for rare-variant association testing, and the approaches can vary in both the null hypotheses being tested and what data can be incorporated (for example, quantitative traits and pedigree information). Unlike collapsing analyses, rare-variant burden methods[23–25] aggregate the information found within a defined genetic region into a summary dose variable. In weighted burden tests[26], variants are additionally weighted according to their frequency or functional impact. Adaptive burden tests[25,27–31] try to account for bidirectional effects by selecting appropriate weights. Variance component (kernel) tests such as C-alpha[32] or SKAT[33] also allow for bidirectional effects, but they are underpowered compared to collapsing or burden tests if many variants are causal and/or if effects are mostly unidirectional within a gene, which, on the basis of existing evidence, seems to be the case for several diseases. Therefore, omnibus tests such as SKAT-O[34] use a combination of burden and variance component tests, which helps for settings with limited prior knowledge of the underlying disease architecture.

The details underlying these approaches and the comparisons between different rare-variant association tests have been reported in great detail elsewhere[35]. Here we focus on the application of rare-variant collapsing analyses as a conceptually simple representative that has proven successful in various settings. However, many of the recommendations also apply to burden and other rare-variant association tests.

Controlling artefactual signals

Sequencing-based cohort analyses are sensitive to variability in the sequencing data of individual samples. Adopting a single common bioinformatic pipeline for the entire test cohort is crucial for reducing the variability

Allelic heterogeneity
The presence of different pathogenic variants in the same gene or at the same chromosome locus that all lead to the same or to very similar phenotypes.

Causal allele
A functional allele that increases disease risk.

Haploinsufficient disease genes
Disease-associated genes for which a single functional copy is insufficient to maintain normal function. Therefore, loss-of-function alleles are pathogenic even when heterozygous.

Background variation
Usually benign variants in the general population that are unconnected to the disease.

Bidirectional effects
Effects within a given gene, wherein some variants increase risk of disease, while others reduce risk.

that comes from different secondary analysis pipelines. In addition to ensuring a single common bioinformatic pipeline, various sequencing quality properties can introduce bias if their distributions differ significantly between the case and control samples. The overall goal is to design a study to minimize heterogeneity between the case and control samples for quality metrics such as the average capture-region read depth, the capture specificity from exome sequencing, the DNA specimen source (blood, saliva, amplified lymphoblastoid cell line (LCL), and so forth), the transition/transversion ratio (Ti/Tv), the proportion of the protein-coding sequence that has adequate coverage per index sample, and other bioinformatic pipeline metrics that could be correlated with elevated rates of exome-wide qualifying variants occurring preferentially in one group.

In this section, we summarize some of the important considerations for reducing bias and thus optimizing disease risk signal detection as part of our common workflow (FIG. 1).

Sample selection. The starting point of all case–control studies is identifying an appropriate control population (FIG. 1Aa,Ab). It is important to select individuals as controls who have not been ascertained for the trait or for a well-known comorbidity with the trait being studied. Additionally, both the cases and controls need to undergo basic quality control (QC) checks, and samples that, for instance, show high contamination rates, low capture region specificity or low coverage need to be removed. Subsequently, as was typical in the common-variant studies, multiple cohort-refining steps are necessary, to minimize contamination within the test cohort.

In population-based sequencing efforts such as the UK Biobank, case–control selection is often performed within the same study cohort. Therefore, cases and controls are typically sequenced together and processed jointly, decreasing the importance of some of the QC procedures mentioned here.

An imbalance of genders among cases and controls can cause problems in collapsing analyses for genes on the sex chromosomes. This becomes an issue particularly in recessive models, where male samples are automatically treated as hemizygous for X-chromosomal variants. Ways to deal with this problem are to analyse the sex chromosomes separately for males and females, to assess matched male/female ratios[14] or to use a tool specifically designed for the analysis of chromosome X, such as XWAS[36].

Another biological confounder that has become surprisingly relevant in rare-variant studies is the effect of an individual's age at DNA specimen collection. Age-associated clonal haematopoiesis caused by acquired mutations in myeloid cancer-associated genes[37–39] can contribute to the observation of inflated rates of qualifying variants among an elderly sampled population. Although the variant allele ratios of these will often be lower than the expected germline rate of 50%, they can remain higher than the lower thresholds commonly adopted during variant filtering. Interestingly, we previously published a collapsing analysis in which the case population was ascertained for amyotrophic lateral sclerosis (ALS), generally a late-onset adult disorder, with a mean range of onset among the WES samples being 57.1 ± 13.0 years of age[13]. Two known genes affected by age-associated clonal haematopoiesis[37,40], *DNMT3A* and *ASXL1*, showed elevated rates of qualifying variants among ALS cases. While it might be associated with increased ALS risk, to date this association has not been established, and it seems likely that this observation was primarily driven by a sample age-related biological signal. This age-related effect is expected to be a practical problem for a small subset of genes involved in haematological cancers, and as such could be accommodated by surveillance of the top collapsing analysis signals, without necessitating strict restrictions to the cohort design. However, due to this effect, for any study exploring the application of collapsing or burden analyses among the somatic mutation class, it is critical to appropriately correct for age at sample collection.

Sample pruning. Since genetic relatedness can distort a variant's contribution to the test statistic, it is necessary to eliminate detectable genetic relatedness from the test cohort in a standard collapsing analysis (FIG. 1Ba). In addition to double-counting true risk alleles among related cases, close relatives within the test also risk cancelling out qualifying variants in ultra-rare test settings, which can then result in deflated qualifying-variant rates. In this sample-pruning step, similar to common-variant GWAS study designs, a single representative from each genetically related pair should be removed from the test cohort until the cohort reflects a collection of unrelated index samples. If a cohort contains a larger number of related individuals, methods can be used that are designed to incorporate family information. These methods typically use measures of genetic similarity to account for relatedness information via linear mixed models[41,42].

Similarly, the inclusion of samples coming from under-represented genetic ancestries can result in inflated rates of 'rare' qualifying variants[43] (FIG. 1Bb). Especially if the case-versus-control composition is greatly imbalanced for genetic ancestry and not all are well-represented in internal and/or external reference cohorts, then the group with genetic ancestries that are under-represented in publicly available reference cohorts will report higher exome-wide rates of the qualifying variants, due to the reduced information that we have about their true frequencies within that under-represented ancestral population[43]. Index samples from consanguineous populations or bottlenecked populations can have the opposite effect, presenting with lower rates of exome-wide ultra-rare qualifying variants due to a reduction of genetic diversity in those populations.

Interestingly, however, if all the samples are well-represented in internal and/or external reference cohorts, the shift in focus to rare variants, especially when focusing on ultra-rare variants, reduces the conventional GWAS concern of population stratification. The simple reason for this is that with large enough sample sizes, very little structure underlies very rare variants,

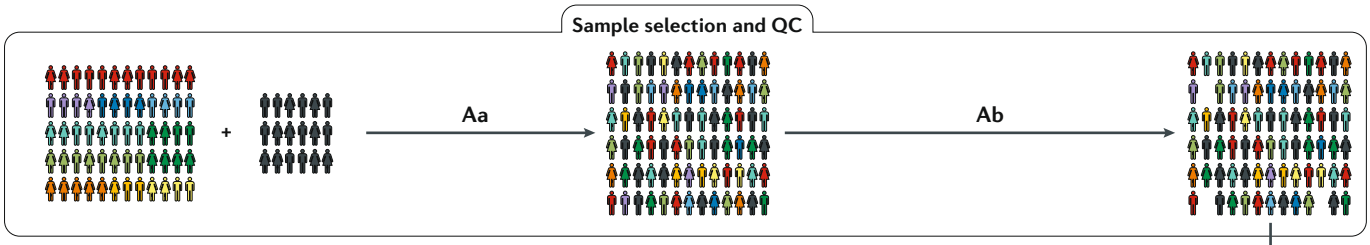
Transition/transversion ratio
(Ti/Tv). Ratio of the number of transitions (interchanges of two-ring purines (A to G or vice versa) or of one-ring pyrimidines (C to T or vice versa)) to the number of transversions (interchanges of purine for pyrimidine bases).

Index samples
Individual samples or patients who are the focus of a study.

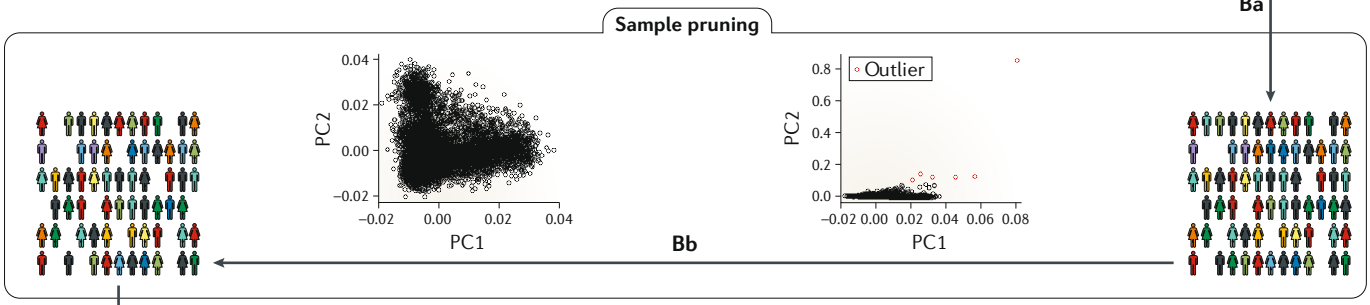
Consanguineous populations
Populations in which marriages between people who are second cousins or closer are common.

Bottlenecked populations
Populations that have gone through a severe and abrupt reduction in their number of individuals, which often leads to reduced genetic diversity.

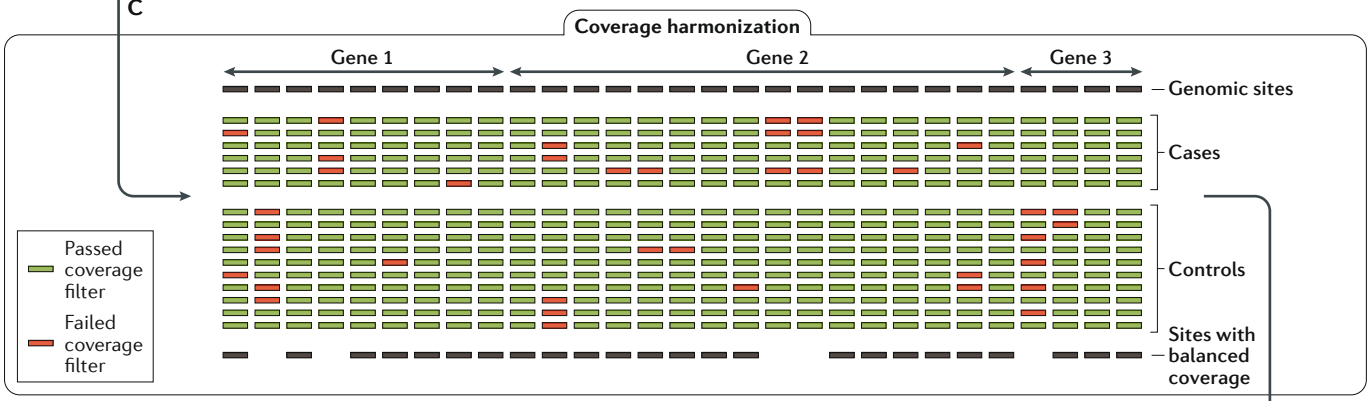
Population stratification
Also known as population structure. Presence of a difference in allele frequencies due to systematic differences in ancestry between cases and controls.



Ba



C



D

QV selection and collapsing

Gene	Cases			Controls			
	S1	S2	S3	S4	S5	S6	S7
G1	0	0	0	1	0	1	1
G2	1	0	1	0	0	0	0
G3	0	1	0	0	1	0	1
G4	0	1	0	0	0	0	0
G5	1	0	0	1	0	0	0
…							…
Gn	0	1	0	0	1	0	0

E

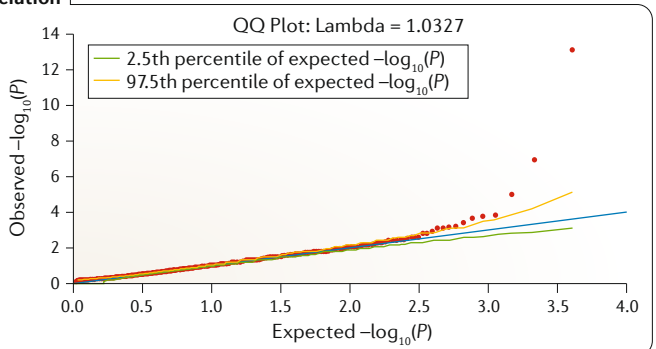
Variant	Gene	Sample	Variant quality filter	Variant annotation filter	External MAF filter	Internal MAF filter
V1	G1	S1	✓	✓	✓	✓
V1	G1	S2	✗	✓	✓	✓
V3	G2	S3	✓	✓	✗	✓
V4	G3	S3	✓	✓	✓	✓
V5	G3	S4	✓	✗	✓	✗
V6	G4	S5	✓	✓	✓	✓
…						…
Vn	Gn	S1	✓	✓	✓	✓

Fa

Test for association

Gene	With QV		Without QV		*P* value
	Cases	Controls	Cases	Controls	
G1	4	0	116	2,500	4e-06
G2	0	0	120	2,500	1
G3	1	1	119	2,499	0.090
G4	2	0	118	2,500	0.002
G5	1	0	119	2,500	0.046
…					…
Gn	0	1	120	2,499	1

Fb



◀ Fig. 1 | **Outline of the standard collapsing analysis approach.** First, cases and matching controls are selected (part **Aa**), and the same sample-level quality control (QC) is performed for cases and controls (part **Ab**). The sample-pruning level comprises relatedness pruning (part **Ba**) and the removal of population outliers based on principal components (PCs) (part **Bb**). All the remaining samples are used to perform coverage harmonization (part **C**), in which sites and therefore variants that show coverage differences between cases and controls are pruned. All remaining variants are used for qualifying variant (QV) selection (part **D**), in which various filters, including internal and external minor allele frequency (MAF) filters, are applied. The selected QVs are used to build the gene-by-individual collapsing matrix (part **E**), which indicates the presence of at least one QV. Finally, each gene is tested for an association between QV status and the phenotype of interest (part **Fa**), and the results can be evaluated by means of a quantile–quantile (QQ) plot (part **Fb**).

and effectively there is no structure underlying ultra-rare variants — variants that are unique to the index sample in the test[44].

Approximately a quarter of all possible protein-coding variants result in a synonymous change. Although some synonymous variants may not be benign, taking the collection of synonymous variants found exome-wide, it is appropriate to consider synonymous variants as a nearly neutral class of genetic variation with respect to many diseases. These synonymous variants then become an important class of genetic variation with respect to good experimental design. This class of mostly neutral or nearly neutral variants provide an important opportunity to evaluate each sample's exome-wide tally of qualifying variants under the null hypothesis of no significant difference between cases and controls. This exercise allows for the identification and exclusion of significant sample outliers in the test cohort distribution of exome-wide synonymous variants under the minor allele frequency (MAF) of interest. Ensuring no significant deviation between the case and control distributions for sample-level exome-wide tallies of synonymous variants is one effective way to confirm that the qualifying variant comparisons are based on a relatively homogeneous cohort. Seeing a significant difference suggests that an underlying bias might subsequently influence the test statistics, and resolving this would best be addressed in advance.

Restricting the analysis to a certain population of interest can also be important if different genes are thought to be responsible for a condition in different populations. Therefore, it is often helpful to perform additional tests on population-specific subsets of the test cohort.

Coverage harmonization. A critical element in any rare-variant statistical test is to ensure that all tests are performed on the genomic sequence in which qualifying variants are equally able to be called in cases and controls. When the tested cohort differs in sequencing characteristics between cases and controls, a signal can be detected that is not caused by the disease and therefore must be corrected in order to control inflation. For example, the sequencing depth or exome capture kits used can result in very different expected numbers of qualifying variants between cases and controls, leading to substantial inflation in test statistics. There are numerous ways to control for this bias, including directly modelling read distributions in the association statistics[45]. This approach is computationally complex but has the advantage of being able to address extreme differences in coverage between cases and controls. A far simpler strategy to minimize potential bias due to coverage differences between cases and controls evaluates each protein-coding exome site — regardless of whether it is variant or invariant — and excludes genomic coordinates from the test if they are not equally covered in cases and controls. There are different approaches to removing differences in coverage: bases can be excluded if less than 90% of either the cases or controls have a pre-defined minimal adequate coverage (often defined as ≥10-fold coverage)[46]; if the case and control populations show differing proportions of individuals with enough coverage[13]; or if a binomial test shows that the case/control status and coverage are not independent[47]. Coverage harmonization (FIG. 1C) is a critical step, as sites where cases and controls are highly imbalanced in their ability to call a variant can become susceptible to false-negative calls in the less well-represented group, leading towards an enrichment bias among the better-represented group. This solution also provides researchers with an estimate of the proportion of each gene (or other defined genomic region) that was harmonized and thus suitable for the statistical test.

If cases and controls are sequenced and processed together, such as in the UK Biobank, the risk of systematic differences between cases and controls is usually low, and coverage harmonization may be skipped.

Selection of qualifying variants. In general, collapsing models are designed with particular parameters, to focus the analysis on specific types of qualifying variants. Whereas QC filters are used for all models, other filters, such as the predicted variant effects or population allele frequencies, depend on the specific model in use. In addition to filters based on popular variant-QC scores, such as Phred quality (QUAL), genotype Phred quality (GQ), quality by depth (QD), mapping quality (MQ) and variant quality score log-odds (VQSLOD), it is also important to filter out known artefacts that repeatedly result in false-positive variant calls. Although there are multiple reasons for the occurrence of these artefacts, such as sequencing errors or problems during alignment or variant calling, they are often specific to a certain capture kit, which can lead to a bias if capture kits differ between the cases and controls.

Reliable annotations are needed to focus on those variants that change a protein's function. As the goal of collapsing is not to be a sensitive clinical diagnostic tool, but rather to optimize signal detection by reducing contamination from neutral background variation, analyses usually focus on protein-truncating, canonical splice site, in-frame insertion or deletion (indel), and missense variants. Furthermore, relying on the consensus coding sequence enables investigators to focus on variants found only in transcripts considered to be of high confidence, while eliminating variants that only affect rare isoforms that might not result in functioning proteins. Various bioinformatic tools can predict the likely effect of missense variants on the protein (for example, CADD[48], PolyPhen-2 (REF.[49]), SIFT[50], REVEL[51] and PrimateAI[52]), and others identify possible splice variants masquerading

Phred quality
(QUAL). The Phred-scaled posterior probability that all samples in a call set consist of homozygous reference alleles.

Genotype Phred quality
(GQ). Represents the Phred-scaled confidence that the genotype assignment is correct for a given sample.

Quality by depth
(QD). The Phred quality (QUAL) score normalized by allele depth for a variant.

Mapping quality
(MQ). Estimation of the overall mapping quality of reads supporting a variant call.

Variant quality score log-odds
(VQSLOD). A score, produced by the Genome Analysis Toolkit's variant quality score recalibration, that represents the log-odds ratio of a variant being true versus being false under the trained Gaussian mixture model.

as 'neutral' intronic or synonymous variants (for example, TraP[53] and SpliceAI[54]) or flag low-confidence LOF variants that are known to be enriched for annotation errors (for example, LOFTEE)[55]. Further analyses on several real datasets are needed to determine the appropriate thresholds for all these methods.

The ability to focus analyses on variants with very low population allele frequencies is arguably the most important reason that collapsing analyses now show good power for several different diseases and traits. The internal cohort MAF should be based on the combined test cohort and not just on the control MAF, as the latter will bias the test. For external MAF filtering (for example, based on the Genome Aggregation Database (gnomAD)), the same MAF filter should be applied equally across case and control variants, and no individuals represented in external datasets adopted for MAF filtering should be included in internal sample sets for the case–control comparisons.

It must be noted that using strict filters not only in terms of variant quality, but also based on predictions of deleteriousness and low MAF, may lead to some loss in sensitivity. However, results across a range of diseases have shown that in a non-diagnostic setting, the increase in specificity due to the filtering outweighs the risk of missing some disease-causing variants[14,15,47,56–58].

Only variants passing all filters of a specific model are termed qualifying variants (FIG. 1D) and subsequently used for building the gene-by-individual indicator matrix used for collapsing. In a dominant genetic model, a 0 in this matrix reflects no qualifying variants found in that gene in that individual, and a 1 reflects at least one qualifying variant found in that gene in that individual (FIG. 1E).

Test for association. The final step of the workflow (FIG. 1Fa,Fb) is to use aggregate statistics (for example, Fisher's exact test, logistic regression or linear mixed models)[59] to find associations between genes with qualifying variants and the phenotype of interest. A quantile–quantile (QQ) plot can be used to evaluate the resulting *P* values.

Statistical considerations. Even though gene-based collapsing combines the information of multiple rare variants into a single value per gene, frequently there will be either no cases or no controls with a qualifying variant in a relevant number of genes, leading to sparse data. In addition, often there is an imbalance between the numbers of cases and controls available. Both characteristics limit the choice of statistical test, because methods that rely on asymptotic properties, such as Pearson's chi-squared test or standard logistic regression, cannot be used. Methods such as Fisher's exact test are preferred; however, they do not easily accommodate covariates. Here, providing evidence of homogeneity between the case and control samples is key[14,15]. An alternative option that allows for the addition of covariate information is Firth correction or a biased reduction logistic regression[60–62]. A score-test-based method that estimates the distribution of the test statistic by using the saddle-point approximation has been proposed as an alternative in unbalanced case–control configurations[63]. Permutation-based application of the chosen test statistic can further help increase the robustness of test results. Empirical *P* values can be calculated as the number of permutation tests achieving a smaller *P* value than the observed test, divided by the number of permutation tests performed. Computational limitations have to be considered, since detecting multiplicity-adjusted significance requires a large number of permutations.

Sparse observations can also lead to biased power calculations if asymptotic properties do not hold[64]. Thus, empirical power calculations need to be performed, by simulating the observed sample data under different parameters[15].

Evaluations of results with QQ plots and the inflation factor lambda are also unreliable with very sparse data and unbalanced case–control configurations. The conventional expected *P* value distribution is based on a chi-squared approximation that relies on the assumption that the *P* values are uniformly distributed under the null hypothesis of no significant difference between cases and controls. This is not an appropriate assumption in certain sparse-data and unbalanced case–control configurations. One working solution is to use an empirical (permutation-based) expected probability distribution[14,15] (BOX 1). Despite being more computationally intensive, this method will be more representative of the true null distribution of the test statistic *P* values for a given study configuration. A complementary test to determine whether exome-wide inflation exists is to compare the proportions of exome-wide qualifying variants in test cases and controls. Observing no significant difference suggests that the exome-wide qualifying variant pool is representative of the case–control ratios[14].

Minor allele frequency resolution offered by public reference cohorts. The MAF resolution associated with publicly available reference cohorts has been a key feature of the success of rare-variant analyses. By taking the sum of the ExAC[65]/gnomAD[55], DiscovEHR[66] and Bravo/TOPMed[67] reference cohorts currently available in 2019, we have access to over 250,000 exomes. Thus, the absence of a test cohort variant at a well-covered protein-coding site across these 250,000 exomes is approximately equivalent to an autosomal MAF of <0.0002%. The current version of gnomAD (2.1.1) on its own, without the addition of other databases, enables investigators to identify protein-coding variants present at MAFs as low as ≤0.0009% in European Caucasians, ≤0.004% in South Asians, ≤0.006% in Africans/African Americans and ≤0.006% in East Asians. Achieving resolution at such low MAFs enables investigators to focus on the youngest alleles present within an index sample — something found to be of great importance for the many traits for which purifying selection affects risk alleles. Indeed, a recurring observation in recent studies of complex disorders has been that the strongest signals from collapsing analyses are concentrated among the rarest of variants[14,15,56,68] (FIG. 2).

Due to the MAF resolution accessible by large population reference cohorts, collapsing analyses have demonstrated utility even in populations that were previously

Box 1 | **Permutation-based expected *P* values**

Under sparse data and unbalanced case–control configurations, the conventional expected *P* value distribution used for quantile–quantile (QQ) plots and the inflation factor lambda becomes unreliable. One working solution is to use an empirical (permutation-based) expected probability distribution. To achieve this, for each collapsing model the original case and control labels (part **a** of the figure) are randomly permuted while keeping the rest of the gene-by-sample matrix fixed (part **b** of the figure). The *P* values for all genes are recomputed with Fisher's exact test using the permuted case–control labels to get the confusion matrix. This is repeated 1,000 times and, for each permutation, the *P* values are ordered. The mean of each rank-ordered estimate across the 1,000 permutations (that is, the average first-order statistic, the average second-order statistic and so forth) represents the empirical estimate of the expected ordered *P* values (part **c** of the figure). This empirical-based expected *P* value distribution no longer depends on an assumption that the *P* values are uniformly distributed under the null hypothesis of no significant difference between cases and controls. The negative logarithm of the permutation-based expected distribution relative to the observed ordered statistic is plotted in order to get the permutation-based QQ plot (part **d** of the figure).

a

	S1	S2	…	S100	S101	S102	…	S300	With QV		Without QV		
Phenotype	0	0		0	1	1		1	Cases	Controls	Cases	Controls	*P* value
Gene 1	1	1		1	0	0		0	0	20	100	280	0.0058
Gene 2	0	1		0	1	0		1	10	0	90	300	6.71e-07
…													…
Gene *n*	1	0		0	0	0		0	2	2	98	300	0.06203

b

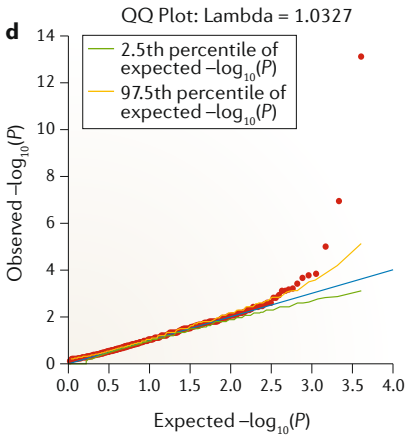
	S1	S2	…	S100	S101	S102	…	S300	With QV		Without QV		
Phenotype	1	0		1	0	1		0	Cases	Controls	Cases	Controls	*P* value
Gene 1	1	1		1	0	0		0	4	6	96	294	0.2761
Gene 2	0	1		0	1	0		1	3	7	97	293	0.7158
…													…
Gene *n*	1	0		0	0	0		0	0	2	100	298	1
…													
	S1	S2	…	S100	S101	S102	…	S300	With QV		Without QV		
Phenotype	1	1		0	0	1		1	Cases	Controls	Cases	Controls	*P* value
Gene 1	1	1		1	0	0		0	3	7	97	293	0.7158
Gene 2	0	1		0	1	0		1	1	9	99	291	0.4624
…													…
Gene *n*	1	0		0	0	0		0	1	1	99	299	0.438

1,000 permutations

c

P1	P2	P3	…	P1000	Mean
0.0001	0.002	0.0004		0.007	0.0009
0.003	0.004	0.0009		0.02	0.002
…					…
1	1	1		1	1

d



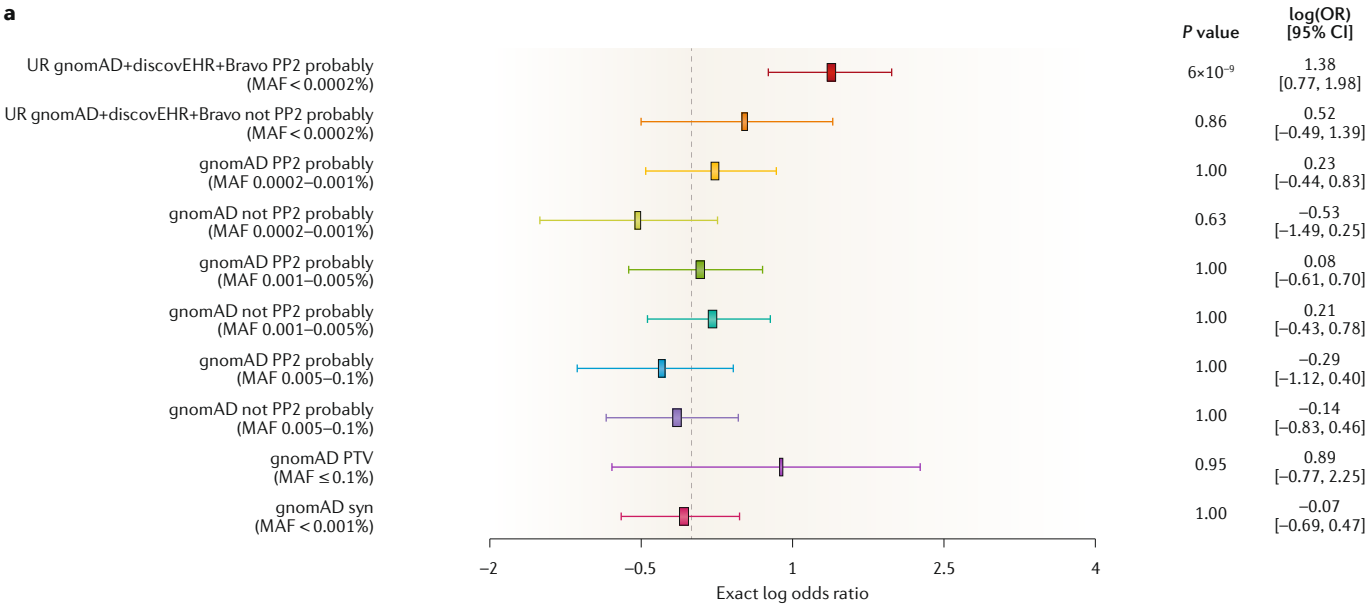
QV, qualifying variant.

considered to be accessible only through trio sequencing. Trio-sequencing approaches transformed our understanding of dominant paediatric Mendelian diseases by pinpointing de novo mutations in the proband that were not constitutively present in parents. However, currently we are often able to identify association signals driven by de novo mutations by screening for qualifying variants that are absent in the large population reference databases available today. This was recently demonstrated in a collapsing analysis study that studied the probands of epileptic encephalopathy trios and was able to successfully implicate multiple genes that had previously been implicated through a trio study design using the same families. Rediscovery of the same signal was enabled by our ability to use large population reference cohorts to reduce the number of exome-wide non-synonymous variants per index sample down to the fewer-than-dozen youngest qualifying variants per sample[44].

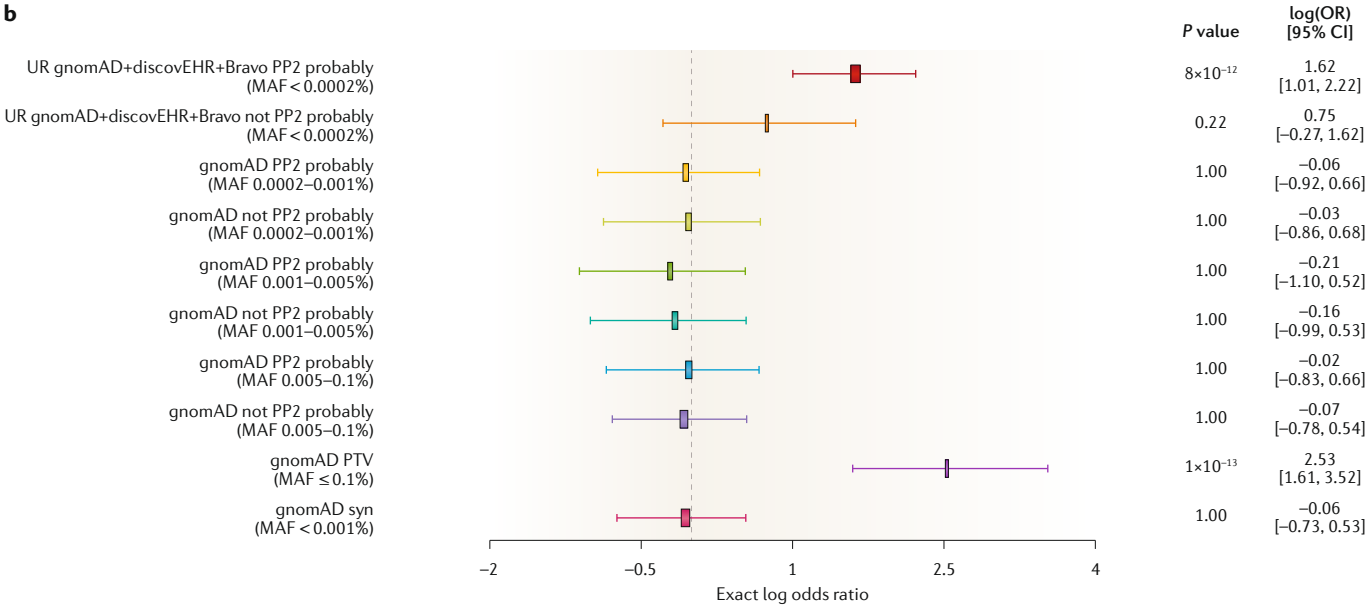
A common question in the application of collapsing studies is whether to use more liberal MAF thresholds. For relatively common diseases, clinically relevant variants may theoretically be present even in multiple copies in databases as large as gnomAD, especially since gnomAD and TOPMed include disease cases as well. However, it is important to emphasize that some common diseases under study are often influenced by a series of strongly acting mutations across tens to hundreds of genes, thus limiting the number of cases that can be explained by individual mutations in individual genes. Moreover, the aim of the collapsing paradigm is not to ensure that every single variant that influences disease risk qualifies for analysis. Instead, the aim is to achieve an effective balance between sensitivity and specificity in the inclusion of clinically relevant variants. Indeed, the success of collapsing analyses has been shown to stem from our ability to minimize the background variation within a gene to allow true genetic-risk alleles that are present due to clinical ascertainment to become prominent in the test. The best way to demonstrate this has been by evaluating the contribution to the detectable risk signal among a collection of known disease genes and across increasing MAF bins (FIG. 2). Thus, the disorders that will benefit most from collapsing analyses are those with high allelic heterogeneity, whereby even if the reference cohorts include a small collection of affected individuals, for the purposes of collapsing tests (that is, gene discovery rather than clinical utility), the increased specificity achieved by strict MAF filters is worth a potential decrease in sensitivity.

Trio sequencing
Procedure in which the index patient and both parents are sequenced in order to identify causative variants in the patient.

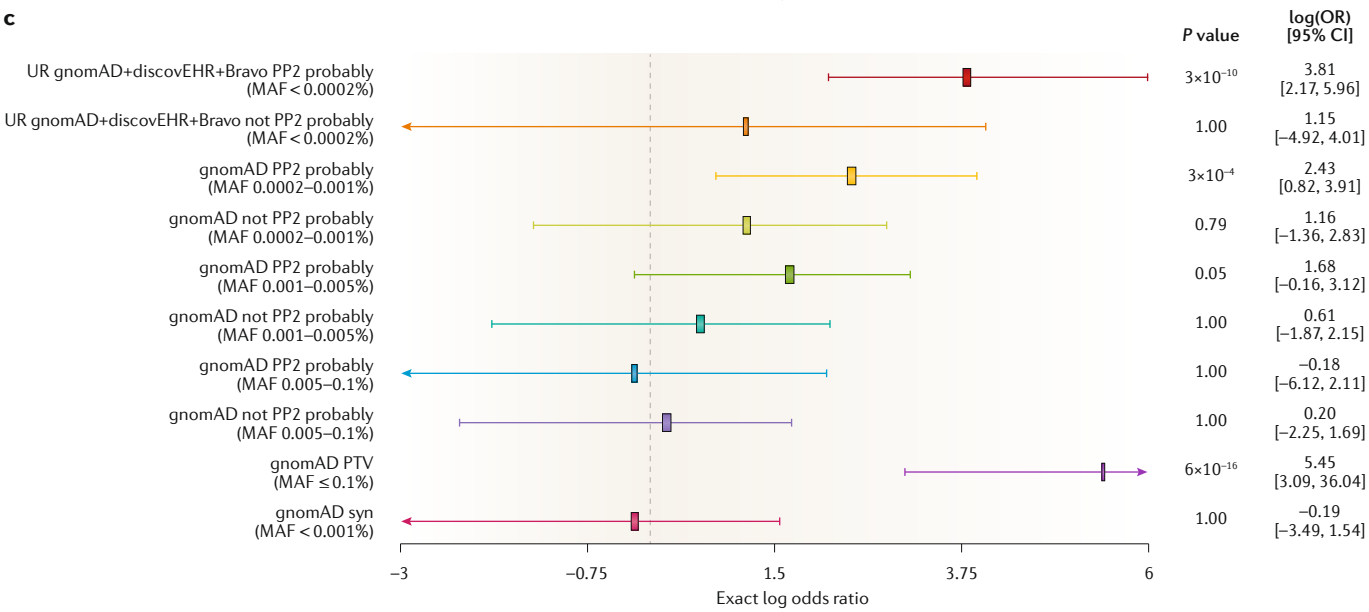
a


P value
log(OR) [95% CI]
UR gnomAD+discovEHR+Bravo PP2 probably (MAF < 0.0002%)
6×10⁻⁹
1.38 [0.77, 1.98]
UR gnomAD+discovEHR+Bravo not PP2 probably (MAF < 0.0002%)
0.86
0.52 [−0.49, 1.39]
gnomAD PP2 probably (MAF 0.0002–0.001%)
1.00
0.23 [−0.44, 0.83]
gnomAD not PP2 probably (MAF 0.0002–0.001%)
0.63
−0.53 [−1.49, 0.25]
gnomAD PP2 probably (MAF 0.001–0.005%)
1.00
0.08 [−0.61, 0.70]
gnomAD not PP2 probably (MAF 0.001–0.005%)
1.00
0.21 [−0.43, 0.78]
gnomAD PP2 probably (MAF 0.005–0.1%)
1.00
−0.29 [−1.12, 0.40]
gnomAD not PP2 probably (MAF 0.005–0.1%)
1.00
−0.14 [−0.83, 0.46]
gnomAD PTV (MAF ≤ 0.1%)
0.95
0.89 [−0.77, 2.25]
gnomAD syn (MAF < 0.001%)
1.00
−0.07 [−0.69, 0.47]
−2
−0.5
1
2.5
4
Exact log odds ratio

b


P value
log(OR) [95% CI]
UR gnomAD+discovEHR+Bravo PP2 probably (MAF < 0.0002%)
8×10⁻¹²
1.62 [1.01, 2.22]
UR gnomAD+discovEHR+Bravo not PP2 probably (MAF < 0.0002%)
0.22
0.75 [−0.27, 1.62]
gnomAD PP2 probably (MAF 0.0002–0.001%)
1.00
−0.06 [−0.92, 0.66]
gnomAD not PP2 probably (MAF 0.0002–0.001%)
1.00
−0.03 [−0.86, 0.68]
gnomAD PP2 probably (MAF 0.001–0.005%)
1.00
−0.21 [−1.10, 0.52]
gnomAD not PP2 probably (MAF 0.001–0.005%)
1.00
−0.16 [−0.99, 0.53]
gnomAD PP2 probably (MAF 0.005–0.1%)
1.00
−0.02 [−0.83, 0.66]
gnomAD not PP2 probably (MAF 0.005–0.1%)
1.00
−0.07 [−0.78, 0.54]
gnomAD PTV (MAF ≤ 0.1%)
1×10⁻¹³
2.53 [1.61, 3.52]
gnomAD syn (MAF < 0.001%)
1.00
−0.06 [−0.73, 0.53]
−2
−0.5
1
2.5
4
Exact log odds ratio

c


P value
log(OR) [95% CI]
UR gnomAD+discovEHR+Bravo PP2 probably (MAF < 0.0002%)
3×10⁻¹⁰
3.81 [2.17, 5.96]
UR gnomAD+discovEHR+Bravo not PP2 probably (MAF < 0.0002%)
1.00
1.15 [−4.92, 4.01]
gnomAD PP2 probably (MAF 0.0002–0.001%)
3×10⁻⁴
2.43 [0.82, 3.91]
gnomAD not PP2 probably (MAF 0.0002–0.001%)
0.79
1.16 [−1.36, 2.83]
gnomAD PP2 probably (MAF 0.001–0.005%)
0.05
1.68 [−0.16, 3.12]
gnomAD not PP2 probably (MAF 0.001–0.005%)
1.00
0.61 [−1.87, 2.15]
gnomAD PP2 probably (MAF 0.005–0.1%)
1.00
−0.18 [−6.12, 2.11]
gnomAD not PP2 probably (MAF 0.005–0.1%)
1.00
0.20 [−2.25, 1.69]
gnomAD PTV (MAF ≤ 0.1%)
6×10⁻¹⁶
5.45 [3.09, 36.04]
gnomAD syn (MAF < 0.001%)
1.00
−0.19 [−3.49, 1.54]
−3
−0.75
1.5
3.75
6
Exact log odds ratio

◀ Fig. 2 | **Characterizing where the disease risk signal resides.** Forest plots of three recently published collapsing analyses: genetic generalized epilepsies (GGE)[15] (part **a**), non-acquired focal epilepsies (NAFE)[15] (part **b**) and idiopathic pulmonary fibrosis (IPF)[14] (part **c**). For each disease group, we used a multivariate logistic regression model to assess the contributions to disease risk from mutually exclusive increasing minor allele frequency (MAF) bins of qualifying variants, similar to the approach described in the original papers[14,15]. The MAF bins for missense variants range from an 'ultra-rare' (UR) definition, based on absence of the variant from a collection of over 250,000 exomes (leveraging variation data from the combination of gnomAD release 2.0, DiscovEHR and Bravo), to an MAF of 0.1%. Protein-truncating variants (PTVs) are included as a single bin with a maximum MAF of 0.1%. As a negative control bin, we used synonymous variants (syn) with a maximum MAF of 0.001%. We split missense variants into those predicted and those not predicted to be 'probably damaging' by PolyPhen-2 (PP2). For the epilepsies (parts **a** and **b**), we adopted the same list of 43 dominant epilepsy genes as in the original paper, and for the IPF cohort (part **c**), we adopted the same three genes (*TERT*, *RTEL1* and *PARN*) as in the original paper. The *x*-axes represent the log-odds ratios, log(OR). All *P* values and accompanying estimates are corrected for the nine tests performed per disease (excluding the tenth, synonymous negative control bin). The strongest disease risk resides in the rarest bin. For all three diseases there is a significant enrichment of UR 'PP2 probably' variants (top red line in each plot). In addition, there is a significant *P* value for MAF bin 0.0002% to 0.001% with PP2 probably for IPF (part **c**), as well as significant enrichment of PTVs for both NAFE (part **b**) and IPF (part **c**).

Beyond dominant single-gene collapsing

Region-based collapsing approach. Different regions within genes can vary in their tolerance to missense variation, and known causal alleles have been shown to preferentially reside in the intolerant sub-regions of disease genes (BOX 2). The accumulation of non-causal variants in more tolerant sub-regions of the same gene reduces the power of gene-based collapsing approaches. Two basic approaches help address this problem: first, collapsing directly on sub-regions (for example, exons or domains) within the genes, which leads to more tests to correct for in the significance threshold, or second, incorporating missense intolerance within genes as an additional filter when selecting qualifying variants. Different regional intolerance measures have recently been proposed: the missense tolerance ratio (MTR)[16], a heuristic tool for measuring the extent of purifying selection acting on missense variants in a given protein-coding window, independent of known genic boundaries; a domain-based MTR[69], which is a complementary approach to subRVIS[70]; a score for missense badness, PolyPhen-2 and constraint (MPC)[19]; constrained coding regions (CCRs)[20], which look for an absence of protein-changing variation over large stretches of coding sequence; and the localized intolerance model using Bayesian regression (LIMBR)[17], which is a hierarchical model that can jointly use genome-wide, genic and sub-region-level information. In the intolerance-informed collapsing approach, regions with intolerance below, for example, the exome-wide 50th percentile can be used for prioritizing variants that are more likely to be clinically relevant, thereby further reducing the noise in the test from background variation. Future studies on real data are needed to show which method and threshold maximize power in a collapsing framework, and whether different diseases need different thresholds, based on factors such as severity, population frequency and the genetic architecture of the disease.

Gene-set collapsing. Especially for studies with smaller sample sizes, noise from background variation can complicate the detection of significant enrichment for qualifying variants within individual genes. To increase signal, genes can be grouped together into biologically informed gene sets and tested for enrichment of qualifying variants among the genes that belong to the set[14,15]. Gene sets can be based on previously associated genes, implicated pathways or other prior knowledge about the disease under investigation. To control for background variation, synonymous qualifying-variant counts within the gene set and qualifying-variant counts in all genes not part of the gene set can be used as covariates in a logistic regression model[15].

Applications to more complex genetic models. Although the collapsing framework so far has been used primarily to identify genes that confer risk due to single dominant-acting mutations, in principle the framework can be applied both to more complex genetic models and to non-coding variation, but with notable complications in its implementation. Perhaps the simplest elaboration of the single-gene-dominant model is consideration of single-gene-recessive models. Despite being straightforward in principle, recessive models are more challenging to implement. One important difference is that internal and external MAF thresholds need to be relaxed, because heterozygous carriers tend to be unaffected, which results in higher MAFs in the general population. A big challenge is caused by the fact that the case–control framework employed in collapsing analyses means that phase is generally unresolved. Therefore, the collapsing framework counts genes with two (or more) qualifying variants as compound heterozygous, when in fact they might be in cis and affect the same copy of the gene, rather than both copies, as is required for recessive inheritance.

Beyond the analysis of single-gene models, the collapsing framework can also be applied to models in which variants in different genes interact to confer risk. Although such digenic or oligogenic models are likely to be important to many complex diseases, the identification of causal genotypes is challenging, due to the high rate of occurrence of qualifying genotypes. A model in which variants in two different, related genes interact to confer risk has been suggested in several conditions, including epilepsy[71,72], Bardet–Biedl syndrome[73,74] and other conditions[75]. If we considered all possible pairwise gene combinations, we would have to perform more than 180 million tests (19,000, choose 2). However, if we take epilepsy as an example, we might be interested in testing for an excess of qualifying genotypes in any two ion channel genes, given the importance of ion channels in epilepsy and evidence of interaction between ion channels in animal models[76]. To illustrate this approach, we used a dataset of 600 cases with genetic generalized epilepsies (GGE) and 2,400 controls from a previously published collapsing analysis[15] and applied lenient internal and external MAF thresholds of 1%. There are around 350 known ion channel genes, amounting to roughly 60,000 possible gene pair combinations. However, in our dataset, for more than 40,000 of those combinations there was no individual with a qualifying variant in both genes. If we further required that more than 5 individuals harbour qualifying variants in both genes, only 1,300 gene pairs were left for testing for

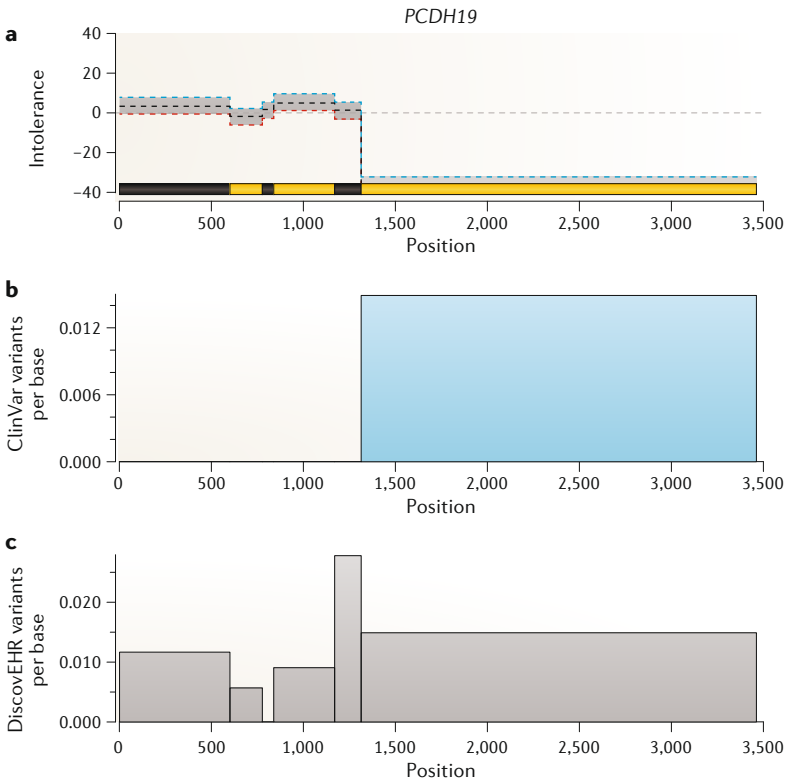
Phase
Defined as alleles that belong to the same parental haplotype and therefore affect the same copy of a gene; variants that are not in phase are on different haplotypes and therefore affect both copies of a gene.

Compound heterozygous
Presence of two different mutant alleles in a particular gene that affect both copies of the gene because they are not in phase.

Box 2 | **Leveraging regional intolerance to missense variation**

To illustrate a possible application of regional missense intolerance, we have plotted the distribution of all pathogenic missense variants in ClinVar[104] compared to the missense variants in DiscovEHR[66] for the epileptic encephalopathy gene *PCDH19* (see the figure). To highlight how missense intolerance scores help distinguish between pathogenic and benign variants, we include scores from the localized intolerance model using Bayesian regression (LIMBR)[17] as an example. LIMBR uses a Bayesian hierarchical model, which facilitates borrowing information across genes, to test for relative depletion in variation at the exon or domain level. In part **a** of the figure, LIMBR intolerance scores are plotted per exon, with 95% credibility in grey across the combined coding positions in all transcripts. The bar strip below that plot indicates the start and end of each exon. The other plots show the densities of variants divided by the number of bases in the exon, matched up at the corresponding genomic positions with the intolerance scores, for ClinVar pathogenic variants (part **b** of the figure) and DiscovEHR control variants (part **c** of the figure).

The plots show that all pathogenic missense variants in ClinVar can be found in the last exon, which has a very low LIMBR score, meaning that it is intolerant to missense variation. By contrast, the variants found in a control population such as DiscovEHR are spread throughout the whole gene and also lie in regions that are very tolerant to missense mutations. Restricting rare-variant collapsing to missense-intolerant regions would increase the signal, because only control variants reflecting background noise would be excluded.



enrichment (see our Digenic analysis tool for the underlying scripts). This illustrates how the challenge of digenic or more complex models might be addressed, by focusing on more biologically informed sets of related genes. Another approach for reducing the search space would be to focus attention on the more intolerant regions of those genes, to remove benign variation.

Tests for non-coding regions. The application of collapsing and related analyses to non-coding regulatory variation is complicated by two primary challenges. First, the tools for recognizing functional variants are currently far more effective for protein-coding genes than for regulatory regions. Second, multiple lines of evidence suggest that single regulatory variants are less likely to have strong effects on gene expression than do protein-coding variants, suggesting that a true burden framework would be more important for regulatory than for coding variation. To effectively apply collapsing and burden analyses to regulatory variation, we would need not only to identify the appropriate regulatory regions within which to aggregate signal, but also to determine the appropriate ways to do the aggregation. Although no formal frameworks have been explored, there are metrics capable of quantifying the importance of non-coding regions, including but not limited to conservation (for example, GERP++[77] and ncGERP[70]), ensemble prediction scores (for example, GWAVA[78] and CADD[48]) and regional human-lineage intolerance of non-coding sequence variation (for example, Orion[79] and ncRVIS[70]). Recently the enhancer domain score (EDS)[80] has been introduced, which can be used independently of and complementary to other metrics of intolerance. The EDS relies on the fact that the number of evolutionarily conserved bases in a gene's enhancers reflects the pathogenicity of protein-disrupting variants in this gene. Because it has been shown that, especially for genes with large enhancer domains, multiple variants in non-coding regions are often necessary to change the expression of the gene, burden tests that take the number of qualifying variants per individual into account may be more successful in this application. Although results from family-based studies have shown that de novo mutations in non-coding regions do contribute to diseases such as autism[81], non-coding variants are currently not routinely included in collapsing-like rare-variant association studies. However, as methods for predicting the functional consequences of non-coding variants evolve, including this class of variant will eventually increase the power of rare-variant association methods.

Combining data. There are multiple ways of leveraging data from different sequencing projects of the same disease. If cases but no controls are sequenced for one of the projects, all cases can be combined and jointly compared to the controls. The same is true if the cases and controls are not well-matched in terms of ancestry, as combining the cohorts could lead to a better balance. Otherwise, the data can also be analysed separately and combined via meta-analysis. The simplest approach would be to directly combine *P* values across studies by using Fisher's[82] or Stouffer's *Z* score[83] method. However, these approaches are not well-powered, compared to joint analysis or more complex methods[84]. Some tools have been developed specifically for the meta-analysis of rare-variant associations, such as MASS[85,86], RAREMETAL[87,88] and MetaSKAT[89]. All these methods use score statistics instead of *P* values and can be used for different types of rare-variant association tests.

Application to complex diseases

The list of diseases for which collapsing analyses have been effective in securely implicating causal genes and also pinpointing individual causal alleles with high confidence continues to grow. Examples range from

application in early-onset paediatric conditions, such as epileptic encephalopathies[44], sudden unexplained death in epilepsy[90] and congenital kidney malformations[91], to more common complex conditions, such as ALS[13,69,92,93], Alzheimer disease[47,94,95], schizophrenia[68,96], epilepsies[15], idiopathic pulmonary fibrosis (IPF)[14] and myocardial infarction[97]. The numbers of cases and controls used range from fewer than 300 cases and 4,000 controls, for IPF[14], to 7,000 cases and 13,000 controls, for Alzheimer disease[47], which illustrates the applicability of collapsing analyses over a wide range of sample sizes.

In addition, gene-based collapsing can also be applied to population-based cohorts, as has been demonstrated recently in an analysis on the UK Biobank data[58]. The authors used electronic health records to obtain thousands of phenotypes for more than 40,000 individuals and looked for gene-based rare-variant associations using the collapsing framework. Furthermore, they performed a successful replication study on over a thousand of the phenotypes, using a separate cohort from a different medical record system. This analysis was the first to look for rare-variant associations in thousands of phenotypes across two large cohorts[58].

Although collapsing analyses can achieve remarkable diagnostic yield (FIG. 2), this analytical framework is not intended to be a replacement for a thorough clinical genetic evaluation. A collapsing analysis is simply the application of a specific set of conditions or rules to a combined case-and-control population to identify where signals of case enrichment exist. As such, if the clinical variant interpretation guidelines are pre-defined, then what the collapsing analysis framework ultimately offers is the ability to perform an objective evaluation of the rates at which such pre-defined qualifying variants occur in a case collection, in comparison to the rates we find them in a sampling of individuals who were not ascertained for the trait of interest (controls). In addition, however, collapsing analyses also provide an opportunity to identify the specific classes of variants for which the detectable signal is found most enriched. This can lead to a critical and objective understanding of the genetic architecture that contributes directly to increased disease risk. In FIG. 2, we see a clear illustration that for three conditions, the missense variant risk signal is most enriched among the ultra-rare missense class of variants — that is, missense variants absent from among over 250,000 available reference control samples. A similar effect has also been shown for protein-truncating variants (PTVs) across multiple phenotypes, particularly psychiatric disorders[56], although that study also highlighted that a role for ultra-rare variation could only be detected for a subset of the diverse disease types analysed. Once a gene is securely implicated via collapsing analysis, it is appropriate to reassess all test cases using more liberal QC, variant effect and MAF thresholds for the new gene(s) under a clinical interpretation paradigm[98].

The future of collapsing analyses

As we have outlined here, collapsing analyses have proven highly effective in identifying genes across a range of disorders in which rare variants in single genes confer substantial risk. We have specified several considerations for the effective application of this framework, and we expect it will continue to implicate new disease genes and help evaluate the contributions of known genes. As highlighted above, we also see considerable opportunity to extend this framework in numerous important directions. We and others have shown that different regions within genes can vary in their tolerance to missense variation, and that known causal alleles preferentially reside in the intolerant sub-regions of disease genes[16–20]. Incorporating missense intolerance within genes as an additional qualifying-variant criterion specifically acting upon missense variants could help further reduce the noise in the test from background variation and thus improve the prioritization of the truly pathogenic missense variants that preferentially affect missense-intolerant regions of a gene, as illustrated in BOX 2.

Similar approaches are also being developed to quantify intolerance to genetic variation among stretches of human non-coding sequence[48,70,77–80,99]. Quantifying and defining the intolerant non-coding boundaries is not trivial, but once available, it will provide valuable information for WGS collapsing applications. Continuous innovation in these non-coding methodologies will further optimize signal detection and lead to better-powered whole-genome collapsing analyses.

Perhaps the single greatest challenge facing rare-variant analyses is the issue of scalability. Sample sizes are increasing, accompanied by corresponding increases in their storage and computational requirements. Large-scale sequencing efforts such as the UK Biobank[100,101] or TOPMed[67] offer the possibility of analysing a large variety of phenotypes[58,102]. However, parallelization and computationally efficient implementations are needed in order to leverage all the available data. Fortunately, the cloud provides a highly flexible solution — operating under a compute-time rent model — to eliminate the need to continuously scale up local institutional clusters. Companies and academic units are already creating the necessary solutions to help address the needs for large-scale population-based rare-variant studies. Many studies commonly generate variants using a joint-calling step across all cohort samples, especially population-based cohort studies such as the UK Biobank or TOPMed, in which cases and controls are sequenced together. However, for projects that combine multiple cohorts that were not sequenced together and in which controls might be re-used for several cohorts, the cost and time required make joint-calling for each analysis rather impractical. This might become even more important in the future, when UK Biobank and TOPMed samples are used as controls for cases sequenced separately. We have successfully used a single-sample haplotype-calling strategy to combine independently generated variants for each sample[14,15]. By using depth of coverage from the alignment step as a proxy for inferring the reference allele at positions without variant calls in a sample, we can create a genotype matrix across all samples, with minimal computation. This strategy only requires single-sample variant call format (VCF) files and a pileup of the depth of coverage from the alignments, which can be collected and stored with a minimal storage footprint. Following the generation of a genotype matrix across a

Diagnostic yield
Rate of discovered diagnostic variants within a collection of cases being tested.

cohort, the statistical analysis can also be computationally intensive. Parallel-processing frameworks have been developed to address the speed and efficiency of these analyses[103]. We anticipate a very urgent need to standardize the generation and storage of these large-scale datasets in a format that is optimized for integration across multiple studies. Such standardization will allow us to maximize the use of these datasets in order to adequately power studies and discover associations with modest effect sizes.

Published online: 11 October 2019

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
2. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
3. Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
4. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425 (2010).
5. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).
6. Need, A. C. et al. Clinical application of exome sequencing in undiagnosed genetic conditions. *J. Med. Genet.* **49**, 353–361 (2012).
7. Zhu, X. et al. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet. Med.* **17**, 774–781 (2015).
8. Yang, Y. et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870–1879 (2014).
9. Appenzeller, S. et al. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am. J. Hum. Genet.* **95**, 360–370 (2014).
10. Homsy, J. et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262–1266 (2015).
11. Fitzgerald, T. W. et al. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
12. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Unlocking Mendelian disease using exome sequencing. *Genome Biol.* **12**, 228 (2011).
13. Cirulli, E. T. et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* **347**, 1436–1441 (2015).
Cirulli et al. present one of the first implementations of collapsing analyses in a case–control study of a complex disease, introducing the qualifying-variant framework, coverage correction and other methodological details.
14. Petrovski, S. et al. An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **196**, 82–93 (2017).
15. Allen, A. S. et al. Ultra-rare genetic variation in common epilepsies: a case–control sequencing study. *Lancet Neurol.* **16**, 135–143 (2017).
This study provides an implementation of collapsing analyses in epilepsy that explicitly evaluates signal as a function of MAF, showing that the association signal observed in epilepsy genes is concentrated amongst the rarest variants.
16. Traynelis, J. et al. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res.* **27**, 1715–1729 (2017).
17. Hayeck, T. J. et al. Improved pathogenic variant localization via a hierarchical model of sub-regional intolerance. *Am. J. Hum. Genet.* **104**, 299–309 (2019).
This research uses a hierarchical model for regional intolerance that can jointly use genome-wide, genic and sub-region-level information.
18. Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. & Goldstein, D. B. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* **17**, 9 (2016).
This paper and reference 19 introduce regional intolerance scoring.
19. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at *bioRxiv* https://doi.org/10.1101/148353 (2017).
20. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
21. Guo, M. H. et al. Determinants of power in gene-based burden testing for monogenic disorders. *Am. J. Hum. Genet.* **99**, 527–539 (2016).
22. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
23. Asimit, J. L., Day-Williams, A. G., Morris, A. P. & Zeggini, E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum. Hered.* **73**, 84–94 (2012).
24. Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
25. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
This study presents one of the early burden-testing methods for rare variants.
26. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLOS Genet.* **5**, e1000384 (2009).
27. Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* **70**, 42–54 (2010).
28. Liu, D. J. & Leal, S. M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLOS Genet.* **6**, e1001156 (2010).
29. Ionita-Laza, I., Buxbaum, J. D., Laird, N. M. & Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLOS Genet.* **7**, e1001289 (2011).
30. Hoffmann, T. J., Marini, N. J. & Witte, J. S. Comprehensive approach to analyzing rare genetic variants. *PLOS ONE* **5**, e13584 (2010).
31. Price, A. L. et al. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
32. Neale, B. M. et al. Testing for an unusual distribution of rare variants. *PLOS Genet.* **7**, e1001322 (2011).
33. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
This study introduces a score-based variance-component test (SKAT) that allows for modelling bidirectional effects.
34. Lee, S. et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
SKAT-O is a unified test that combines burden tests with the non-burden sequence kernel association test.
35. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
36. Gao, F. et al. XWAS: a software toolset for genetic data analysis and association studies of the x chromosome. *J. Hered.* **106**, 666–671 (2015).
37. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
38. Buscarlet, M. et al. DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood* **130**, 753–762 (2017).
39. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
40. Carlston, C. M. et al. Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for bohring-opitz syndrome. *Hum. Mutat.* **38**, 517–523 (2017).
41. Lippert, C. et al. Fast linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
42. Oualkacha, K. et al. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet. Epidemiol.* **37**, 366–376 (2013).
43. Petrovski, S. & Goldstein, D. B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* **17**, 157 (2016).
This study emphasizes the importance of the geographic ancestry of controls for both collapsing analyses and identifying pathogenic mutations in patients.
44. Zhu, X. et al. A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations. *PLOS Genet.* **13**, e1007104 (2017).
This report illustrates that collapsing analyses in a case–control design focused on the rarest variants can pick up the same variants as analyses of de novo mutations using trios.
45. Hu, Y.-J., Liao, P., Johnston, H. R., Allen, A. S. & Satten, G. A. Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. *PLOS Genet.* **12**, e1006040 (2016).
46. Guo, M. H., Plummer, L., Chan, Y.-M., Hirschhorn, J. N. & Lippincott, M. F. Burden testing of rare variants identified through exome sequencing via publicly available control data. *Am. J. Hum. Genet.* **103**, 522–534 (2018).
47. Raghavan, N. S. et al. Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. *Ann. Clin. Transl. Neurol.* **5**, 832–842 (2018).
48. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
49. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
50. Sim, N.-L. et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
51. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99, 877–885 (2016).
52. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
This study describes a deep neural network trained on hundreds of thousands of common variants from population sequencing of six non-human primate species that can identify pathogenic variants.
53. Gelfman, S. et al. Annotating pathogenic non-coding variants in genic regions. *Nat. Commun.* **8**, 236 (2017).
54. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
55. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Preprint at *bioRxiv* https://doi.org/10.1101/531210 (2019).
56. Ganna, A. et al. Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum. *Am. J. Hum. Genet.* **102**, 1204–1211 (2018).
Ganna et al. show that across multiple phenotypes, rarer PTVs are on average more deleterious, with the strongest signal coming from ultra-rare variants.
57. Cameron-Christie, S. et al. Exome-based rare-variant analyses in CKD. *J. Am. Soc. Nephrol.* **30**, 1109–1122 (2019).
58. Cirulli, E. T. et al. Genome-wide rare variant analysis for thousands of phenotypes in 54,000 exomes.

Preprint at *bioRxiv* https://doi.org/10.1101/692368 (2019).
This analysis is the first to look for rare-variant associations in thousands of phenotypes across two large cohorts, including the UK Biobank data.
59. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
60. Wang, X. Firth logistic regression for rare variant association tests. *Front. Genet.* **5**, 187 (2014).
61. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).
62. Heinze, G. & Puhr, R. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Stat. Med.* **29**, 770–777 (2010).
63. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).
64. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).
65. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
66. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
67. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Preprint at *bioRxiv* https://doi.org/10.1101/563866 (2019).
68. Genovese, G. et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
69. Gelfman, S. et al. A new approach for rare variation collapsing on functional protein domains implicates specific genic regions in ALS. *Genome Res.* **29**, 809–818 (2019).
70. Petrovski, S. et al. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLOS Genet.* **11**, e1005492 (2015).
71. Baulac, S. et al. Evidence for digenic inheritance in a family with both febrile convulsions and temporal lobe epilepsy implicating chromosomes 18qter and 1q25-q31. *Ann. Neurol.* **49**, 786–792 (2001).
72. Ito, M. et al. Phenotypes and genotypes in epilepsy with febrile seizures plus. *Epilepsy Res.* **70**, 199–205 (2006).
73. Fauser, S., Munz, M. & Besch, D. Further support for digenic inheritance in Bardet–Biedl syndrome. *J. Med. Genet.* **40**, e104 (2003).
74. Katsanis, N. et al. Triallelic inheritance in Bardet–Biedl syndrome, a Mendelian recessive disorder. *Science* **293**, 2256–2259 (2001).
75. Schäffer, A. A. Digenic inheritance in medical genetics. *J. Med. Genet.* **50**, 641–652 (2013).
76. Glasscock, E., Qian, J., Yoo, J. W. & Noebels, J. L. Masking epilepsy by combining two epilepsy genes. *Nat. Neurosci.* **10**, 1554–1558 (2007).
77. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput. Biol.* **6**, e1001025 (2010).
78. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
79. Gussow, A. B. et al. Orion: detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLOS ONE* 12, e0181604 (2017).
80. Wang, X. & Goldstein, D. B. Enhancer redundancy predicts gene pathogenicity and informs complex disease gene discovery. Preprint at *bioRxiv* https://doi.org/10.1101/459123 (2018).
81. An, J.-Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
82. Fisher, R. A. *Statistical Methods for Research Workers* (Oliver and Boyd, 1932).
83. Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. & Williams Jr, R. M. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II)* Vol. 1 (Princeton Univ. Press, 1949).
84. Liu, L. et al. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLOS Genet.* **9**, e1003443 (2013).
85. Tang, Z.-Z. & Lin, D.-Y. MASS: meta-analysis of score statistics for sequencing studies. *Bioinformatics* **29**, 1803–1805 (2013).
86. Tang, Z.-Z. & Lin, D.-Y. Meta-analysis of sequencing studies with heterogeneous genetic associations. *Genet. Epidemiol.* **38**, 389–401 (2014).
87. Feng, S., Liu, D., Zhan, X., Wing, M. K. & Abecasis, G. R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828–2829 (2014).
88. Liu, D. J. et al. Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
89. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **93**, 42–53 (2013).
90. Bagnall, R. D. et al. Exome-based analysis of cardiac arrhythmia, respiratory control, and epilepsy genes in sudden unexpected death in epilepsy. *Ann. Neurol.* **79**, 522–534 (2016).
91. Sanna-Cherchi, S. et al. Exome-wide association study identifies greb1l mutations in congenital kidney malformations. *Am. J. Hum. Genet.* **101**, 789–802 (2017).
92. Freischmidt, A. et al. Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia. *Nat. Neurosci.* **18**, 631–636 (2015).
93. Farhan, S. M. K. et al. Enrichment of rare protein truncating variants in amyotrophic lateral sclerosis patients. Preprint at *bioRxiv* https://doi.org/10.1101/307835 (2018).
94. Christophersen, I. E. et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **49**, 946–952 (2017).
95. Bellenguez, C. et al. Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. *Neurobiol. Aging* **59**, 220.e1–220.e9 (2017).
96. Singh, T. et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat. Genet.* **49**, 1167–1173 (2017).
97. Do, R. et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102–106 (2015).
98. Groopman, E. E. et al. Diagnostic utility of exome sequencing for kidney disease. *N. Engl. J. Med.* **380**, 142–151 (2019).
99. Telenti, A. et al. Deep sequencing of 10,000 human genomes. *Proc. Natl Acad. Sci. USA.* **113**, 11901–11906 (2016).
100. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015).
101. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
102. Van Hout, C. V. et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. Preprint at *bioRxiv* https://doi.org/10.1101/572347 (2019).
103. Zhang, D. et al. SEQSpark: a complete analysis tool for large-scale rare variant association studies using whole-genome and exome sequence data. *Am. J. Hum. Genet.* **101**, 115–122 (2017).
104. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).

Acknowledgements
The authors thank T. Hayeck for creating the figure in Box 2.

Author contributions
G.P., S.P. and J.H. researched data for the article. G.P., S.P. and D.B.G. wrote the article. G.P., S.P., J.H., A.S.A. and D.B.G. reviewed/edited the manuscript before submission. All authors contributed to discussing the content of the article.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Genetics thanks S. Lee, X. Lin and B. Neale for their contribution to the peer review of this work.

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

RELATED LINKS
Bravo/TOPMed: https://bravo.sph.umich.edu/freeze5/hg38/
CADD: https://cadd.gs.washington.edu/
CCR: https://www.rebrand.ly/ccrregions
ExAC: http://exac.broadinstitute.org/
Digenic analysis tool: https://github.com/igm-team/Digenic
gnomAD: http://gnomad.broadinstitute.org/
LOFTEE: https://github.com/konradjk/loftee
MTR: http://mtr-viewer.mdhs.unimelb.edu.au/
PolyPhen-2: http://genetics.bwh.harvard.edu/pph2/
PrimateAI: https://github.com/lllumina/PrimateAI
SIFT: https://sift.bii.a-star.edu.sg/
SpliceAI: https://github.com/illumina/SpliceAI
subRVIS: http://www.subrvis.org/
TraP: http://trap-score.org/
UK Biobank: https://www.ukbiobank.ac.uk/