



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/245455/>

Version: Published Version

Article:

Tithi, S.S., Cooper-Knock, J., Benatar, M. et al. (2026) CoCoRV-nf: a powerful and cost-effective tool for rare variant analysis leveraging external biobank sequence data identified new candidate predisposition genes in amyotrophic lateral sclerosis and neuroblastoma. Human Molecular Genetics, 35 (17). ddag076. ISSN: 0964-6906

<https://doi.org/10.1093/hmg/ddag076>

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial (CC BY-NC) licence. This licence allows you to remix, tweak, and build upon this work non-commercially, and any new works must also acknowledge the authors and be non-commercial. You don't have to license any derivative works on the same terms. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

CoCoRV-nf: a powerful and cost-effective tool for rare variant analysis leveraging external biobank sequence data identified new candidate predisposition genes in amyotrophic lateral sclerosis and neuroblastoma

Saima Sultana Tithi¹, Johnathan Cooper-Knock², Michael Benatar³, Joanne Wu³, J. Paul Taylor¹, Gang Wu^{4,*}, Wenan Chen^{5,*}

¹Department of Cell & Molecular Biology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, United States

²Sheffield Institute for Translational Neuroscience, University of Sheffield, Western Bank Sheffield S10 2TN, United Kingdom

³Department of Neurology and ALS Center, University of Miami Miller School of Medicine, 1600 NW 10th Avenue, Miami, FL 33136, United States

⁴Center for Applied Bioinformatics, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, United States

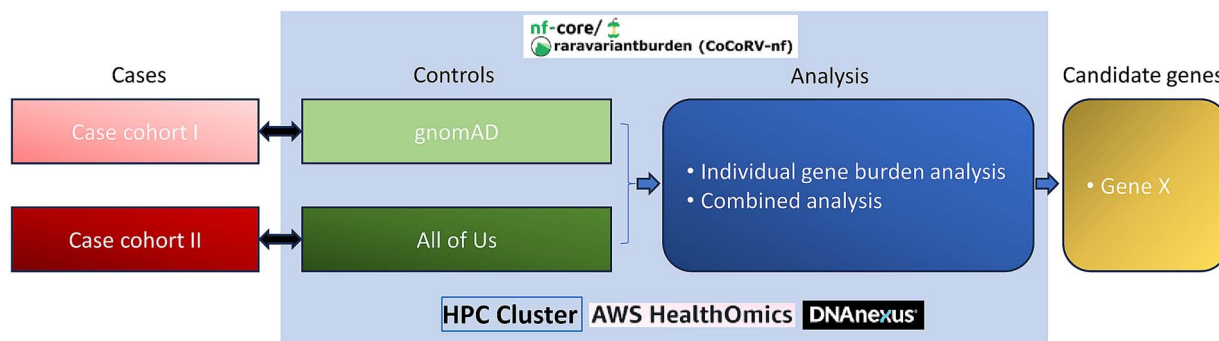
⁵Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester 55905, United States

*Corresponding authors. Gang Wu, Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN, USA. E-mail: gang.wu@stjude.org; Wenan Chen, Department of Health Sciences Research, Mayo Clinic, Rochester, USA. E-mail: chen.wenan@mayo.edu

Abstract

Although sequencing costs have steadily decreased with advances in technology, they remain high for large scale studies. The design of traditional individual-disease sequencing studies is either case only or cases with relatively few controls, resulting in potential loss of statistical power for discovery of disease associated genes. Here we show that for a given number of sequenced cases, a large control sample size is critical to maximize power for rare variant burden analysis. Furthermore, we have developed an end-to-end workflow based tool (CoCoRV-nf) to facilitate the use of external biobank sequence resources as controls. The modules include consistent variant QC, variant annotation, ancestry population prediction, and gene based burden analysis using summary genotype information, and combined analysis from multiple independent results. The tool supports exomes and genomes from gnomAD and All of Us as controls with preprocessed datasets. We apply the tool in two rare neurological diseases: amyotrophic lateral sclerosis and neuroblastoma. For each disease, two case cohorts are paired with gnomAD and All of Us data, respectively, followed by a combined analysis. Not only did we recapture known genes, but also, we identified new candidate genes for both diseases. By leveraging multiple large external biobank sequence data, we demonstrate the feasibility of using our tool to maximize statistical power to identify new disease predisposition genes.

Graphical abstract



Keywords rare variant burden analysis, predisposition genes, external controls, biobank sequence resources

Received: March 14, 2026. Revised: July 13, 2026. Accepted: July 20, 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Introduction

Whole exome sequencing (WES) and whole genome sequencing (WGS) provide an unprecedented opportunity to identify rare germline pathogenic variants that predispose to diseases and traits, which cannot be tagged by previous genotyping arrays [1, 2]. In order to discover predisposition genes with rare pathogenic germline variants, the sample sizes of both cases and controls are critical to ensuring adequate statistical power. Due to high cost, for studies of individual diseases, sequencing is often performed only for cases or cases together with a relatively small number of controls. This design substantially reduces the statistical power in gene based rare variant association analysis. Therefore, strategic investment in building large biobank level datasets has been initiated, including gnomAD [3, 4], UK Biobank [5], and All of Us (AoU) [6]. However, the analyses of the sequenced case cohorts outside of these biobank datasets remain challenging. To address the challenge, we previously developed a tool consistent summary counts based rare variant burden test (CoCoRV) [7] to facilitate association analyses including variant QC, genetic ancestry matching, and summary genotype count based gene burden test, leveraging the large sample size of already sequenced controls. Nevertheless, several important problems remain to be addressed, together with new opportunities for improved analysis. First, it is not clear which studies will benefit the most from the large control sizes, considering that the gain of statistical power will saturate when the control sample size reaches a certain threshold [8]. Second, the previously developed version of CoCoRV could not be applied to sex chromosomes in rare variant analysis, which may result in missing important sex chromosome linked causal genes. Third, given the paradigm shift of using cloud based platforms for data storage and analysis, tools that are applicable both in local high performance computing (HPC) clusters and in cloud platforms, are preferred. Fourth, it is important to leverage multiple large sequencing biobanks for combined analysis when multiple independent case cohorts are available. To address the above issues, we have developed a tool called CoCoRV-nf, where nf stands for the use of Nextflow in executing multiple modules. We demonstrate its effectiveness through application to two rare neurological diseases for identifying predisposition genes. For each disease, we collected two independent case cohorts, then we paired each of them with gnomAD exomes and AoU as controls, respectively. Then we performed a combined analysis to aggregate the results.

Results

CoCoRV-nf provides an end-to-end workflow for gene based burden analysis

CoCoRV-nf uses Nextflow [9] to organize different analysis components, supporting both local HPC clusters and cloud environments (Fig. 1a, Fig. S1). The main features of CoCoRV-nf are summarized in Table 1. CoCoRV-nf provides inference of ancestry groups using gnomAD's random forest based classifier, important for pairing cases with controls of the same continental ancestries. It supports both ANNOVAR and VEP for variant annotations (Fig. 1b). Another main feature is accommodation of large biobank controls in the analysis: gnomAD v2.1 WES & WGS (GRCh37), gnomAD v4.1 WES & WGS (GRCh38), and AoU exome regions (v8, GRCh38), all re-annotated using ANNOVAR and VEP, and stored in the GDS format for efficient storage and retrieval. Because of multiple available controls, multiple cases

Table 1 Summary of main features in CoCoRV-nf.

Features	Details
Automatic inference of ancestry groups Support ANNOVAR and VEP for variant annotations	gnomAD's random forest based ancestry group classifier is applied to case samples Either one or both can be specified for variant annotation, which can be used for defining variant groups in analysis
Sex stratified analysis supporting sex chromosomes Multiple resources as controls	The number of alleles for PAR regions and non-PAR regions are separately calculated for males and females gnomAD v2.1 exomes & genomes (GRCh37) gnomAD v4.1 exomes & genomes (GRCh38) All of Us v8 WGS exome regions (GRCh38) Combine independent CoCoRV analysis results for improved power
Combined analysis	

cohorts can be paired with multiple controls, respectively, and then a combined analysis to further increase sample sizes.

Large control size can substantially improve statistical power for rare variant analysis

We performed statistical power analysis of genetic associations on control sizes considering three different settings: high penetrance ultra rare variants, high risk rare variants, moderate risk common variants. Results show distinct patterns among the three settings (Fig. 2, Table S1). First, the case sample size is the most important factor influencing statistical power for rare variant analysis when the control sample size is large enough or at the saturation point. When the number of cases is fixed, as expected, power continues to increase as the control sample size increases, however, there are differences when the power reaches a plateau. The setting of high penetrance ultra rare variants benefits the most from the large number of controls. For example, with 1200 cases, the power is less than 0.1 with 2000 controls, but it reaches around 0.5 with 10 000 controls, ~0.8 with 50 000 controls, and over 0.9 with 100 000 controls. The power increase at the sample size of gnomAD v4.1 WES is substantial compared with that at the sample size of gnomAD v2.1 WES for 600 and 1200 cases (Fig. 2a). Under the setting of high risk rare variants power still plateaus at a large control size, e.g. 50 000 or 100 000, except almost no power for 200 cases (Fig. 2b). For the setting of moderate risk common variants, power plateaus around 5000 or 10 000 controls. In summary, the simulation results show that, unlike common variants, a large control sample size is critical to achieving sufficient statistical power for rare variant analysis, which underscores the advantages of applying CoCoRV-nf with large external biobank data as controls for rare variant analysis.

Combined analysis of PGB-ALS and ANSWER-ALS identified two new ALS candidate predisposition genes with validation evidence

For ALS, 5–10% is familial (with identifiable ALS predisposition genes in ~70%) [10], the rest are sporadic ALS (with identifiable ALS predisposition genes in ~7%). Therefore, in our analysis, we stratified ALS cases into sporadic ALS (n = 407 in PGB, n = 345 in ANSWER-ALS) and familial ALS (n = 63 in PGB, n = 368 in ANSWER-ALS) and paired with gnomAD v4.1 exomes and AoU v8, respectively (Fig. 3).

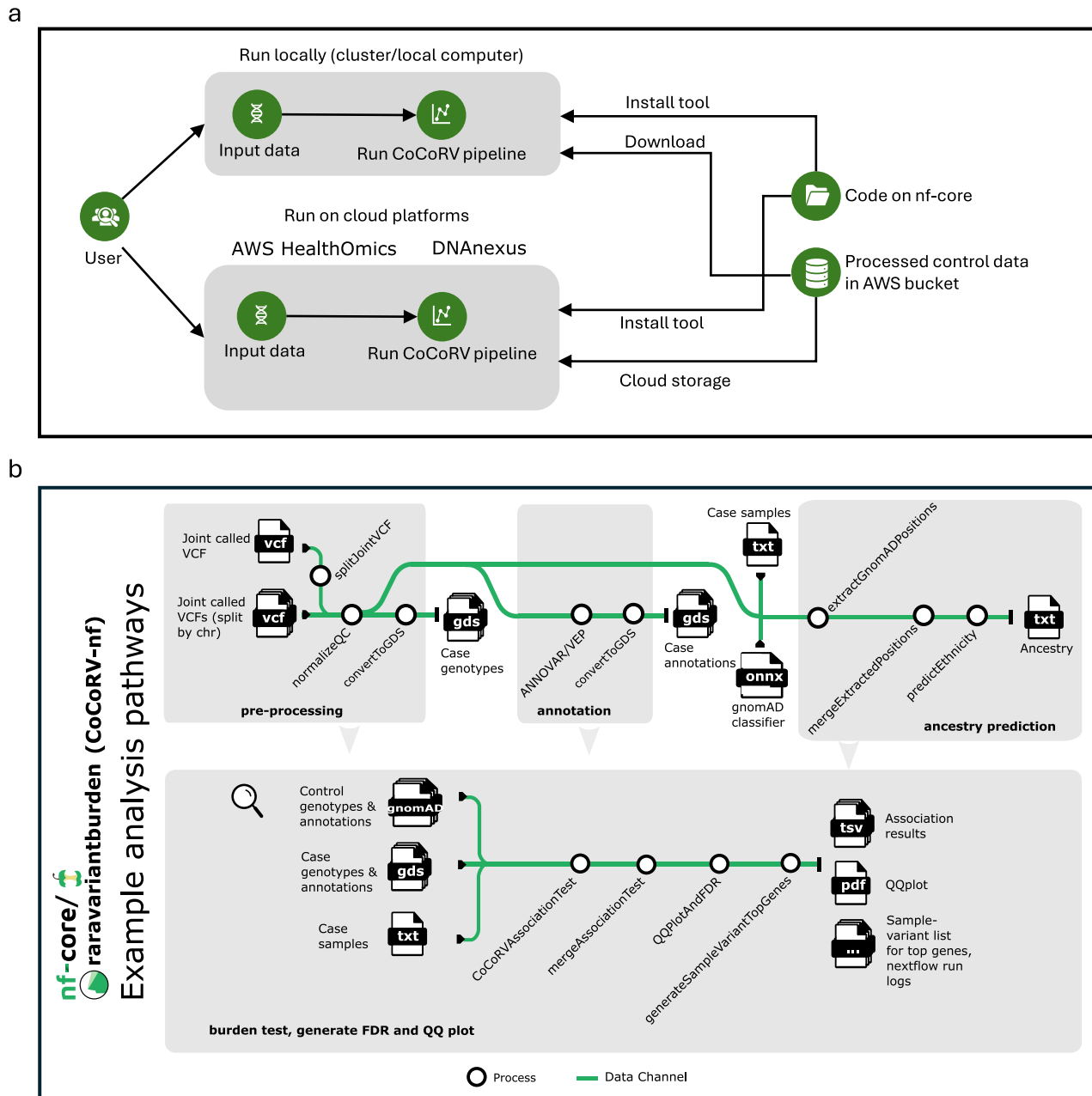


Figure 1 Schematic overview of CoCoRV-nf. a. the CoCoRV-nf pipeline can be run locally such as local clusters or a local computer, it can also be run on cloud platforms including AWS and DNAnexus. We provide processed control data in an AWS bucket including annotated variants saved in gds format. b. Main workflow modules of CoCoRV pipeline implemented on nf-core. It takes the jointly called VCF files after systematic variant QC such as GATK VQSR as input, then performs variant normalization and further QC, annotation, sample ancestry prediction and then gene based rare variant burden analysis.

There is no systematic inflation for either individual cohort results or combined analysis (Fig. 4). For both sporadic and familial ALS, we identified the well-known ALS gene *SOD1* as the top gene from the combined analysis (Table 2). *SOD1* is also the top gene for three of four individual case-control analysis except in PGB-ALS sporadic analysis (Table S2 and S3). Known ALS genes *NEK1* and *TARDBP* are also significant ($P = 2E-4$ and $P = 4E-4$, respectively) and ranked within top 5 in the combined analysis result of sporadic ALS. Using FDR cutoff 0.2, we found two new genes: *SLC52A2* ($FDR_{combined} = 0.03$) from sporadic ALS analysis and *BCCIP* ($FDR_{combined} = 0.07$) from familial

ALS analysis. IGV based read visualization showed good quality of the variants included in the burden test (Figs S2 and S3).

The *SLC52A2* gene encodes riboflavin transporters and recessive mutations are known to cause Brown-Vialetto-Van Laere syndrome 2 (BVVLS2) [11], a neurodegenerative disorder. In the ALS cohort, all mutations are heterozygous (Table S2d). The *BCCIP* gene (BRCA2 and CDKN1A-interacting protein) is crucial in maintaining genomic stability and cell cycle progression. For both genes, there are both loss of function variants and deleterious missense variants in mutation carriers.

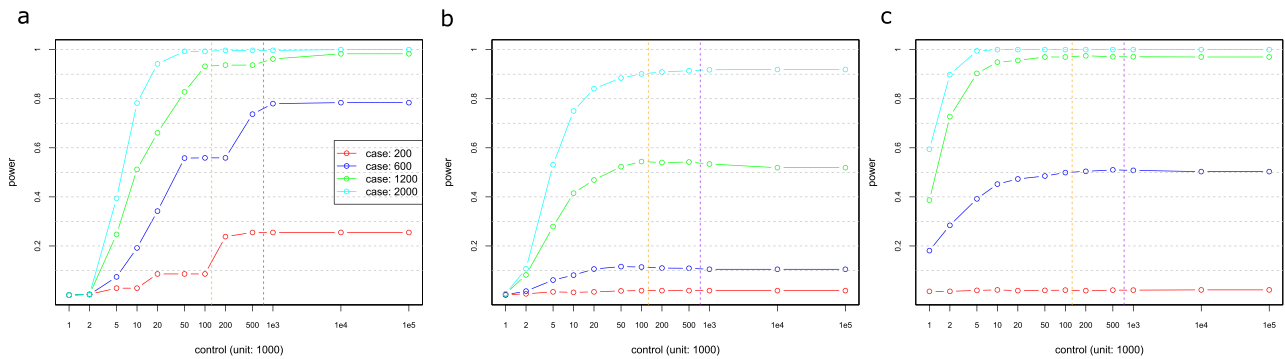


Figure 2 Power simulation results with different control sizes under different effect sizes and allele frequency settings. a. Mendelian high penetrance ultra rare variants; b. high risk rare variants; c. moderate risk common variants. The left vertical line corresponds to the sample size of gnomAD V2 WES samples, and the right vertical line corresponds to the sample size of gnomAD v4 WES samples.

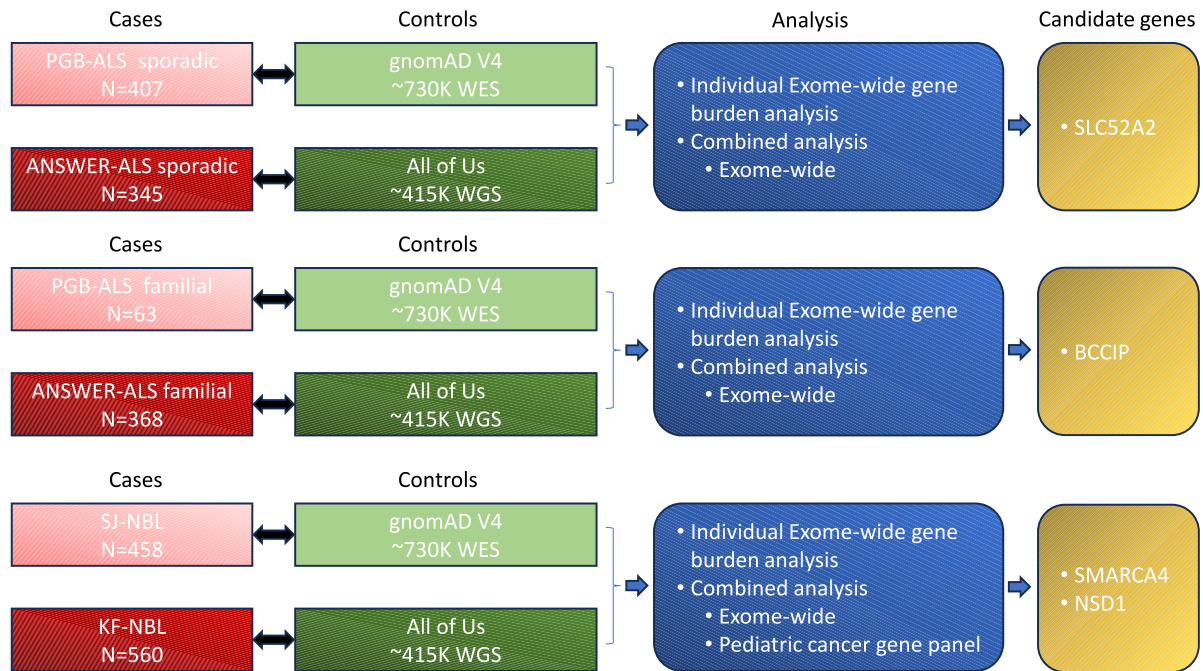


Figure 3 Analysis design of ALS and NBL cohorts.

Table 2 Top genes of combined analysis results of sporadic ALS and familial ALS.

		PGB sALS				ANSWER sALS				Combined analysis			
Gene	P	OR	CI _{Lower}	CI _{Upper}	P	OR	CI _{Lower}	CI _{Upper}	P	OR	CI _{Lower}	CI _{Upper}	FDR
Sporadic													
SOD1	1.28E-01	7.34	0.18	41.76	3.39E-11	66.98	26.12	144.03	2.71E-10	33.19	14.14	66.64	1.39E-06
SLC52A2	3.24E-04	8.81	2.83	20.84	8.43E-03	7.41	1.51	22.04	9.03E-06	8.22	3.53	16.37	2.62E-02
NEK1	1.06E-02	6.79	1.39	20.11	7.48E-03	7.75	1.58	23.05	2.32E-04	7.24	2.64	15.9	4.79E-01
TARDBP	1.00E+00	0	0	335.71	1.57E-04	122.89	13.68	516.49	4.32E-04	70.34	8.18	273.1	6.81E-01
Familial													
SOD1	7.14E-09	210.68	54.16	593.52	2.99E-09	54.4	19.33	124.06	4.71E-16	78.68	36.55	150.94	2.02E-12
BCCIP	2.27E-02	44.69	1.1	261.79	4.40E-04	21.42	4.33	64.74	2.75E-05	24.69	6.61	65.04	7.29E-02

FDR: false discovery rate; P: P value; OR: odds ratio; CI_{Lower}, CI_{Upper}: lower and upper bound of 95% confidence interval of odds ratios.

We analyzed gene expression in 99 patients with ALS spectrum motor neuron disease and 11 non-neurological controls from NYGC ALS Consortium cohort. There are significant differences between

cases and controls for *SOD1* (P value = $1.6E-5$, Fig. S4). Among two new genes, *BCCIP* shows significantly different expressions between ALS patients and controls Fig. 4d and h.

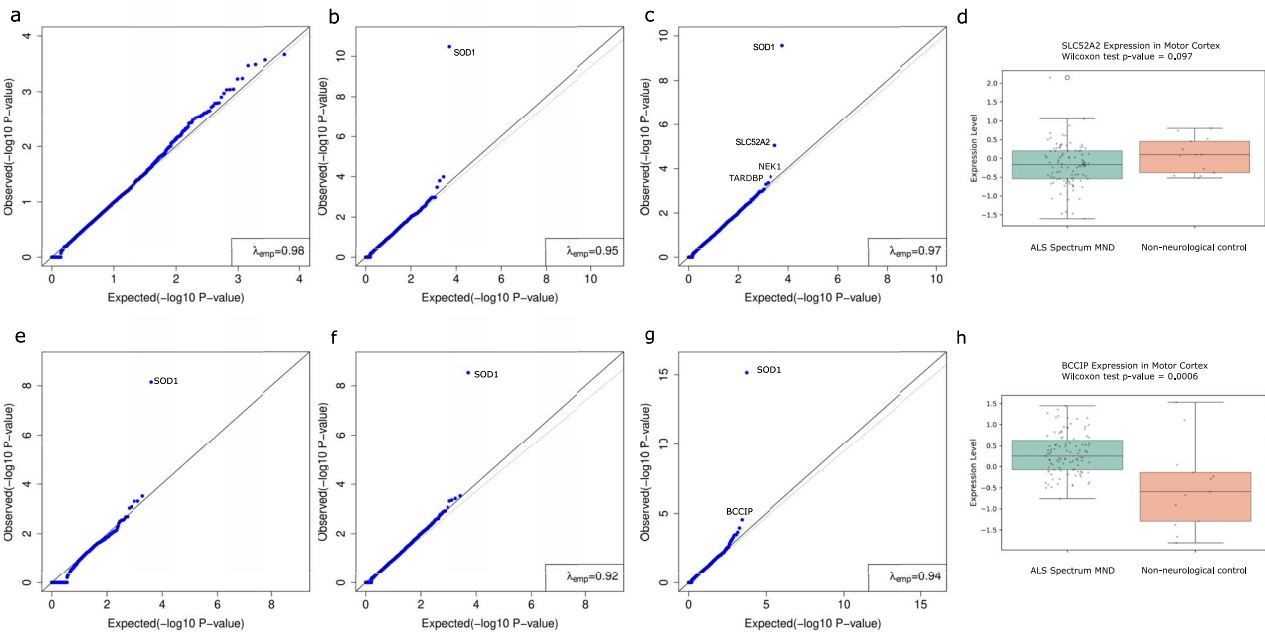


Figure 4 QQ plots of gene based rare variant burden test and gene expression analysis for sporadic ALS (sALS) and familial ALS (fALS). a. PGB sALS versus gnomAD v4.1 exome; b. ANSWER sALS versus all of us; c. combined sALS analysis results using the combined method 'cat'; d. gene expression of SLC52A2 between ALS Spectrum motor neuron disease (MND) patients and non-neurological controls using RNASeq data from NYGC ALS consortium. e. PGB fALS versus gnomAD v4.1 exome; f. ANSWER fALS versus all of us; g. combined fALS analysis results using the combined method 'cat'; h. gene expression of BCCIP between ALS Spectrum motor neuron disease (MND) patients and non-neurological controls using RNASeq data from NYGC ALS consortium. The inflation factor λ was not estimated in e because the number of P values < 1 is less than 20 after removing the top 5% quantile of observed $-\log_{10} P$ values. The dark diagonal line represents the expected line for perfect alignment between observed and expected $-\log_{10} P$ -values, the gray dashed line represents the fitted line of the points where the slope corresponds to the empirical inflation factor λ_{emp} .

Table 3 Published gene based analysis results of the two new ALS genes.

	Project MinE		Chinese ALS		Harmonized WES/WGS ALS	
	P	OR	P	OR ^a	P	OR
<i>SLC52A2</i>	0.026	1.68	0.053	>1	0.620	1.09
<i>BCCIP</i>	0.836	0.94	0.21	>1	0.036	1.50

For Project MineE and Chinese ALS, all variants that passed filtering were used. For the harmonized WES/WGS ALS, high impact + moderate variants were used. Note that the harmonized WES/WGS ALS data sets include both Project MinE and Chinese ALS. ^aSKATO was used, so only the enrichment direction was reported, > 1 means enriched in cases.

Next we sought replication using gene based rare variant analysis result from Project MinE (6538 cases and 2415 controls) [12], Chinese ALS (610 ALS cases, 460 controls) [13], and one large harmonized WES/WGS based ALS analysis (~13 000 cases and ~70 000 controls) which includes both Project MinE and the Chinese ALS data sets (Table 3) [14]. *SLC52A2* is significant ($P = 0.026$, OR = 1.68) from Project MinE, and with border-line significance ($P = 0.053$) in Chinese ALS, considering all variants when combining both loss of function and missense variants. For the large harmonized WES/WGS based analysis, *SLC52A2* is not significant, even though the trend is the same. *BCCIP* is not significant either in Project MinE or the Chinese ALS study, but is significant in the large harmonized ALS analysis ($P = 0.036$, OR = 1.50).

We also performed sex stratified analysis including chromosome X for both familial and sporadic PGB-ALS versus gnomAD exomes. Results are similar to the analysis without sex stratification for autosomal genes. However, we identified the known X-linked ALS

gene *UBQLN2* with borderline significance in PGB-ALS sporadic cases ($P = 0.052$).

Combined analysis of SJ-NBL and KF-NBL identified two new NBL candidate predisposition genes

There is no systematic inflation for either individual analysis or combined analysis (Fig. S5). For analysis using putative pathogenic variants including deleterious missense variants, we first validated that the known predisposition gene *PHOX2B* is significant in SJ-NBL but not in KF-NBL ($P_{SJ-NBL} = 0.003$, $P_{KF-NBL} = 1$, Table S4a), *BARD1* significant in KF-NBL but not in SJ-NBL ($P_{KF-NBL} = 0.001$, $P_{SJ-NBL} = 1$, Table S4b). Then we sought discovery of novel predisposition genes. There were no exome-wide significant genes in the combined analysis with FDR < 0.2 (Table S4, S5). We further restricted our analysis to genes in the 103 predefined pediatric cancer gene list summarized by Christian P Kratz [15] to reduce

Table 4 Top genes of combined analysis results of NBL focusing on LoF variants of pediatric cancer genes.

gene	SJ-NBL				KF-NBL				Combined analysis				
	P	OR	CILower	CIUpper	P	OR	CILower	CIUpper	P	OR	CILower	CIUpper	FDR
SMARCA4	8.38E-04	49.7	5.85	189.28	1.00E+00	0	0	71.96	4.28E-03	21.38	2.54	79.97	0.16
NSD1	1.00E+00	0	0	66.91	2.00E-03	32.1	3.76	123.63	6.79E-03	16.73	2	62.11	0.16
SDHA	1.47E-02	11.07	1.33	40.66	4.12E-01	1.89	0.05	10.68	3.55E-02	4.24	0.87	12.53	0.41
PHOX2B	1.33E-02	79.53	1.89	512.56	1.00E+00	0	0	185.4	3.58E-02	28.55	0.69	174.21	0.41
NF1	1.02E-02	13.53	1.62	49.97	1.00E+00	0	0	13.33	6.97E-02	4.68	0.56	17.09	0.48

FDR: false discovery rate; P: P value; OR: odds ratio; CILower, CIUpper: lower and upper bound of 95% confidence interval of odds ratios.

the number of multiple testings. We found three significant genes with $FDR < 0.2$ when restricting genes to the predefined pediatric cancer genes. For putative pathogenic variants including deleterious missense variants, known gene *PHOX2B* is significant ($FDR_{combined} = 0.10$, $FDR_{SJ-NBL} = 0.01$, $FDR_{KF-NBL} = 1$, Table S6). For LoF variants, there are two new significant genes (Table 4, Table S7a), *SMARCA4* ($FDR_{combined} = 0.16$, $FDR_{SJ-NBL} = 0.01$, $FDR_{KF-NBL} = 1$) and *NSD1* ($FDR_{combined} = 0.16$, $FDR_{SJ-NBL} = 1$, $FDR_{KF-NBL} = 0.06$). *SMARCA4* is associated with multiple cancers such as rhabdoid tumor, there are two mutation carriers in SJ-NBL, one stop-gain and one frameshift (Table S7b).

A recently published study suggested *SMARCA4* as a predisposition gene for NBL and reported 11 neuroblastoma patients with *SMARCA4* mutations, with one stop-gain variant overlapping with our analysis [16]. *NSD1* is associated with Sotos syndrome with elevated risk of cancers, there are two NBL carriers, one stop-gain and one frameshift variants. Even though there is no report of *NSD1* being a predisposition gene before, *NSD1* hypermethylation was a predictor of poor outcome in high-risk neuroblastoma.

Large control sample size can improve the significance level of rare variant analysis

When the mutation frequency is very low, small control size can reduce power because of the uncertainty in estimating risk allele carrier frequencies in controls. This was also reported before in our analysis of acute lymphocytic leukemia using gnomAD v2 [7]. Here we show another example. In the analysis of SJ-NBL, *PHOX2B* is ranked the second when using gnomAD exomes (~700 000 samples) with $P = 0.00028$. We have also tested the association between SJ-NBL and gnomAD genomes (Table S8), which has ~70 000 samples, about one tenth of the sample size of gnomAD exomes, *PHOX2B* is ranked third with a reduced P value 0.0011. Even though in general significance increases with larger control size, there is no guarantee that gnomAD v4.1 exomes will always achieve more significant P-values than gnomAD v4.1 genomes because of intrinsic fluctuations of mutation carriers in data and power plateau for large control sample sizes.

CoCoRV-nf is cost-effective in the cloud

We compared the computational cost of the analysis from different execution environments. We use the analysis of sporadic PGB-ALS versus gnomAD v4.1 exomes as an example (Table 5), which represents the general trend for other analyses. The computing time ranges are similar, ~5–6 hours for the completion of the full workflow, with local

HPC the fastest mainly due to no overhead of node scheduling and activation as in the cloud. The peak memory is also similar, ~12-13GB. The cloud based cost is \$30 using AWS Health Omics and \$50 using DNAnexus. The cost-effective analysis allows investigators to perform efficient analysis and take advantage of the large external sequencing data as controls.

Discussion

We proposed a powerful and cost effective tool to leverage large external biobank data as controls for gene based burden analysis, which supports both local clusters and cloud environments. By applying the tool to two neurological diseases: ALS and NBL, we not only replicated known predisposition genes, but also identified new candidate genes: *SLC52A2* for sporadic ALS, *BCCIP* for familial ALS, and *SMARCA4*, *NSD1* for NBL. We share the full genome-wide analysis results in the supplementary material, which can be used for future meta-analysis. Application of our tool to other diseases, e.g. rare genetic diseases or cancers, will likely result in discovery of more predisposition genes.

One limitation of our analysis result is that we used a lenient FDR threshold 0.2, and we haven't performed independent validation of the candidate genes. In addition, we used published rare variant analysis results for new ALS candidate genes which included all missense variants in the analysis, not the same as the variant group in our ALS analysis which includes missense variants with REVEL score > 0.65. Therefore, these identified genes are only candidate genes so additional independent validation is needed to confirm their associations. Notably, *BCCIP* was significantly enriched in rare variants in the recent study ($p = 0.03$) [14]. Mutation in *SLC52A2* can cause BVVLS2, characterized by early childhood onset of sensorineural deafness, bulbar dysfunction, muscle weakness, and respiratory insufficiency, which shows phenotypic similarity with ALS. However, BVVLS2 is auto-recessive, while in our analyzed patients all rare variants are heterozygous (Table S2d). In addition, *SLC52A2* carriers in PGB-ALS have age of onset at least 48, and at least 56 in ANSWER-ALS except one missing value. Therefore, the likelihood of diagnosis contamination of the known BVVLS2 is low. Recent reports show genes implicated in autosomal recessive neurological diseases such as HSP could confer the risk for ALS [17–19]. When examining the structure location of observed eight variants, all are within or close to the predicted transmembrane domains of *SLC52A2* (Fig. S6). Four out of nine carriers have known pathogenic variants from ClinVar (two p.G306R from two different genetic variants and two p.L339P).

We further examined the normalized gene expression (TPM) of *BCCIP* of motor neurons from patient induced pluripotent stem cells

Table 5 Cost comparison of running the workflow in cluster and clouds.

	AWS Health Omics	DNAnexus	HPC Cluster
Runtime	6 h 31 m	6 h 47 m	5 h 14 m 44 s
CPU & memory allocated (head node)	omics.m.2xlarge 3.4 GHz, 8vCPUs, 32GB RAM	mem2_hdd2_v2_x4 3.59 GHz, 4vCPUs, 16 GB RAM	cluster node, 2.8 GHz, 32GB RAM
Maximal memory used	12.07 GB	12.46 GB	12.99 GB
Disk allocated/used	2400 GB	2000 GB	768 GB
Cost	\$28.05	\$49.27	NA

(iPSC) in ANSWER ALS. Of the three putative pathogenic mutation carriers (Table S3d), the one with a heterozygous frameshift variant in exon 1 showed ~50% of the normal expression (5.72 vs. median value 10.16), with under-representative alternate allele (Fig. S7), suggesting potential nonsense mediate decay. Another carrier with a heterozygous frameshift variant in the last exon (exon 7) showed normal range expression (13.04), likely escaping nonsense mediated decay. However, the functional effect on the truncated protein is not clear. The one with the nonsynonymous variant in exon 3 showed normal expression (10.09) as expected, suggesting that the impact is likely on the function of the normal expressed mutated protein.

Our initial pairing of controls was based on sequencing instruments and historical time. For example, PGB ALS and SJ NBL spanned over 8 years and were mixed with multiple Illumina sequencers, so it's better to pair with gnomAD, which is also a retrospective collection of multiple large scale studies for over 10 years. While ANSWER ALS and KF NBL were from NovaSeq, which is more consistent with AoU. However, we think other pairings with controls should work in general and lead to similar combined results. We further did a sensitivity for sporadic ALS by switching the control counts in the two case-control pairs and regenerated the P-values of individual datasets as well as the combined results (Table S9). The combined results show very similar results with the same Top 3 genes (*SOD1*, *SLC52A2*, and *NEK1*). For the new candidate genes *SLC52A2*, *P*_{add} of *SLC52A2* is 1.05E-5 instead of 1.20E-5, *P*_{cat} of *SLC52A2* is 7.92E-6 instead of 9.03E-6. We checked the QQ plots which also showed well calibration.

When choosing the control dataset, we think the most important factor is the sample size within each ancestry after making sure the reference build matches. For gene based burden analysis, both WES and WGS controls show well calibrated results in general. We have summarized the sample sizes of both gnomAD and AoU by ancestry in Table S10 to assist the decision. Given that the WES samples in gnomAD is often much larger than WGS in general, gnomAD WES is likely the first choice for gnomAD control datasets. Running the analysis using both gnomAD WES and WGS can provide a sensitivity check. In some cases, gnomAD WGS may provide better coverage of exons than gnomAD WES and be more powerful as shown in our previous analysis [20]. AoU v8 has about four times sample sizes of gnomAD v4.1 WES for African American ancestry and about 3.5 times sample size for Admixed American ancestry, so AoU may improve the statistical power for case cohorts from African American or Admixed American ancestry.

There are also limitations of CoCoRV-nf. One common pattern of false association we found is that a single variant observed in multiple samples drives the association signal of the gene. In this case, we recommend careful examinations before claiming statistical significance. Firstly, check whether the alternate allele read frequencies

(VAFs) are away from expected value 0.5 for a heterozygous genotype. If most VAFs are far less than 0.5, e.g. 0.2–0.3, this indicates poor variant quality, and the variant should be excluded. Secondly, perform a sensitivity check using both gnomAD exomes and genomes. One example is the variant chr12-11 062 035-A-ACC, which is a true variant from gnomAD genomes with overall alternate allele frequency (AF) 0.008 but does not exist from gnomAD exomes even though it is inside the exon of gene *TAS2R46*, which can lead to false positive association if not excluded when using gnomAD exomes. Thirdly, check the possibility of founder mutations only present in the cases. Specifically, we need to check the AF of controls from the more granular population where the cases were collected. For example, if the case population is also represented in gnomAD HGDP/1KG subpopulations, we can compare AF of the more granular subpopulation with that of the general continental population to see if AF is substantially higher in the subpopulation.

So far our application of CoCoRV-nf focuses on the gene coding regions. Given the availability of WGS, in principle noncoding regions can also be analyzed either independently or together with coding regions. More investigation is needed into how to best incorporate the noncoding regions to facilitate the analysis. Another direction is gene set/pathway based analysis, which may provide stronger signal than individual genes. Now multiomics data generated for a specific disease/trait is not uncommon, e.g. ANSWER ALS has both WGS and gene expression data from patient iPSC derived motor neurons. How to best integrate multiomics data to prioritize predisposition genes associated with diseases/traits will be another interesting direction.

Materials and methods

Power simulation with different control samples sizes

To illustrate the effect of control sample sizes on statistical power given a fixed sample size of cases, we performed a series of power simulations. We considered three settings: 1) high penetrance ultra rare variants; 2) high risk rare variants; 3) moderate risk common variants. For all simulations, we assumed a dominant model and used population controls, considering that for rare variants often only heterozygotes are observed, and it is often more convenient to view external data sets as population controls which might contain few cases, such as gnomAD. More simulation details are in Supplementary text. Fisher exact test was used for the association test, and the genome-wide significance threshold was set to 2.5×10^{-6} assuming a Bonferroni adjustment for 20 000 genes. We simulated different control sizes for each fixed number of cases: 200, 600, 1200, 2000, and evaluated the empirical power using 1000 replicates.

Predicting ancestry population using gnomAD's random forest classifier

One critical step in association analysis is to account for population structure, e.g. using principal components (PCs) [21] derived from genotype matrix. Given that only summary count information is available from gnomAD, calculating PCs is not possible. However, we can still match the major ancestry groups to account for the ancestry differences. This ancestry group prediction is implemented using the python package hail and gnomAD python package. We apply the classifier to both gnomAD and AoU controls because AoU uses almost the same ancestry groups as gnomAD.

Sex stratified analysis including chromosome X and Y

We extend CoCoRV based analysis to the sex chromosomes X and Y in our tool. The challenge is the different copies of sex chromosomes between males and females. Our main idea is that we not only stratify the data by ancestry, but also by sex, i.e. all individuals are partitioned into each ancestry-sex group. Then we use the available sex-ancestry specific alternate allele counts (AC) and total allele counts (AN) annotation from gnomAD as control counts.

Generation of summary count resource for all of us and gnomAD

We processed AoU v8 samples (~415 K) into a format similar to that in gnomAD. For AoU, due to the data size, currently only variants in coding and neighboring regions were included, and sex specific AC and AN were not available yet. For those with coverage summary statistics, e.g. gnomAD v2.1, v4.1, we generated a new BED file where there are at least 10× coverage for 90% samples for autosomal chromosomes, and 10× coverage for 40% samples for sex chromosomes.

Variant annotation and defining variant groups

Both ANNOVAR [22] and VEP [23] are supported in our tool. Because we pre-annotated the control variants with both ANNOVAR and VEP in gnomAD exomes, genomes, and AoU, only case data need to be annotated in association analysis, which significantly reduces the analysis time. The annotations are then used to define variant groups. Our tool provides a flexible way in defining variant groups by using R's logical and comparison syntax.

Combined analysis of multiple case-control count tables

For each case-control analysis, CoCoRV outputs the $2 \times 2 \times K$ case-control count matrix for K ancestry groups, where for each ancestry group, it is a 2×2 table between disease status (case or control) and risk allele carrier status (carrier or non-carrier). We provide two ways to combine results from multiple case-control results generated by CoCoRV. The first is concatenation (cat): given N result matrices $2 \times 2 \times K_i, i = 1 \dots N$, concatenate all matrices into one single matrix of dimension $2 \times 2 \times \sum_{i=1}^N K_i$. The second is by adding counts of the same ancestry group (add): for each shared ancestry group from multiple analysis results, add the corresponding counts to form a new 2×2 table for the shared ancestry group, resulting a $2 \times 2 \times L$ matrix assuming L distinct ancestry groups among all analysis results. Once the final count matrix is created, the same Cochran-Mantel-Hansel test and FDR calculation accounting for heterogeneous distributions of genes

are used. The method 'add' assumes that each ancestry group from multiple analyses are homogeneous in terms of frequencies of risk allele carriers in cases and controls. Therefore, the method 'cat' is more robust, even though they show similar results in our analysis.

Gene based rare variant burden analysis and combined analysis of two disease cohorts

Two Amyotrophic lateral sclerosis (ALS) cohorts are the CReATe Consortium's Phenotype-Genotype-Biomarker (PGB) study ($n = 470$) [19] and ANSWER-ALS ($n = 713$) [24]. Two Neuroblastoma (NBL) cohorts are St. Jude pediatric cancer genomes study (SJ-NBL, $n = 458$) [25] and Kids First study (KF-NBL, $n = 560$) [26]. To our knowledge, due to the sparsity of neuroblastoma sequencing data, our analysis includes the largest number of cases and controls among gene based burden analysis in identifying neuroblastoma predisposition genes [27, 28]. All four disease cohorts were whole genome sequenced and aligned to GRCh38 using BWA [29]. GATK [30] was used to generate the gVCF files and jointly call the multi-sample VCF files. Variant QC was then performed using VQSR following the GATK best practices. We used PCRelate [31] to perform relatedness analysis to keep unrelated individuals, e.g. removing related individuals up to 3rd-degree. For PGB, two samples out of 472 were excluded due to relatedness/low quality, resulting in 470 samples used in the analysis. In the gene based burden analysis, gnomAD exomes served as controls for PGB and SJ-NBL, and AoU served as controls for ANSWER-ALS and KF-NBL. After each individual case-control analysis, we performed a combined analysis using the summary count information from two analysis results. The main variant group we used in the analysis is ANNOVAR based putative pathogenic group, which includes stop gain, frameshift insertion or deletion, and missense with REVEL > 0.65 . In the analysis of NBL, we also used the loss of function (LoF) variant group: stop gain, frameshift insertion or deletion, and canonical splicing.

Acknowledgements

We thank those instrumental in the creation and provision of the data sets from Gabriella Miller Kids First (GMKF), CReATe consortium, NYGC ALS consortium, TargetALS and ANSWERALS, and ALS Association. We thank Keith Massey from Cure RTD Foundation for allowing us to use the protein structure figure of SLA52A2. We gratefully acknowledge *All of Us* participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data [and/or samples and/or cohort] examined in this study. The authors would like to thank the Genome Aggregation Database (gnomAD) Consortium and the groups that provided data to this resource for making these data available.

Author contributions

Saima Sultana (Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing—original draft, Writing—review & editing), Johnathan Cooper-Knock (Resources, Writing—review & editing), Michael Benatar (Data curation, Resources, Writing—review & editing), Joanne Wu (Data curation, Resources, Writing—review & editing), J. Paul Taylor (Data curation, Resources, Writing—review & editing), Gang Wu (Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources,

Writing—original draft, Writing—review & editing), and Wenan Chen (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *Human Molecular Genetics* online.

Funding

This work was supported by National Institutes of Health [R03OD039975 to WC, GW, SS, U54 NS092091 to GW, SS] and National Cancer Institute [P30 CA021765 to GW, SS].

Data availability

This study has no new data released. The CReATe PGB ALS raw data is available in dbGaP/AnVIL (phs002844, phs003184) and St. Jude Cloud (SJC-DS-1024). ANSWER ALS data was downloaded from NeuroMine (phs003067, <https://data.answerals.org/>) Neuroblastoma raw data were downloaded from St. Jude Cloud (SJC-DS-1001, SJC-DS-1002, SJC-DS-1004, SJC-DS-1005, SJC-DS-1007) and GMKF CAVATICA (<https://www.cavatica.org/>). Processed gnomAD v2.1 and v4.1 in AWS s3 bucket: s3://cocorv-resource-files/. Analysis using AoU controls is not shared due to AoU policy, but can be done through collaboration with authors.

Competing interests

S.S., G.W., W.C., J.W., J.C., J.P.T. report no competing interests. M.B. reports consulting fees for Alauos, Alector, Alexion, Amgen, Annexon, Arrowhead, Biogen, Bristol Myers Squibb, Canopy, Cartesian, CorEvitas, Denali, Eli Lilly, Immunovant, Janssen, Merck, Novartis, Prilenia, Roche, Sanofi, Takeda, UCB, uniQure, and Woolsey.

Code availability

CoCoRV: <https://bitbucket.org/Wenan/cocorv/>
CoCoRV-nf: GitHub code <https://github.com/nf-core/rarevariantburden>

Ethics approval and consent to participate

This study is determined as 'Non-Human Subjects Research Determination' by St. Jude Children's Research Hospital IRB (20–0545, reference number 008769).

References

- Zhang J, Walsh MF, Wu G *et al*. Germline mutations in predisposition genes in Pediatric cancer. *N Engl J Med* 2015;**373**:2336–2346. <https://doi.org/10.1056/NEJMoa1508054>.
- Wilcox N, Tyrer JP, Dennis J *et al*. The contribution of coding variants to the heritability of multiple cancer types using UK biobank whole-exome sequencing data. *Am J Hum Genet* 2025;**112**:903–912. <https://doi.org/10.1016/j.ajhg.2025.02.013>.
- Karczewski KJ, Francioli LC, Tiao G *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–443. <https://doi.org/10.1038/s41586-020-2308-7>.
- Chen S, Francioli LC, Goodrich JK *et al*. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 2024;**625**:92–100. <https://doi.org/10.1038/s41586-023-06045-0>.
- Consortium, U.K.B.W.-G.S. Whole-genome sequencing of 490,640 UK biobank participants. *Nature* 2025;**645**:692–701. <https://doi.org/10.1038/s41586-025-09272-9>.
- All of Us Research Program Genomics, I. Genomic data in the all of us research program. *Nature* 2024;**627**:340–346. <https://doi.org/10.1038/s41586-023-06957-x>.
- Chen W, Wang S, Tithi SS *et al*. A rare variant analysis framework using public genotype summary counts to prioritize disease-predisposition genes. *Nat Commun* 2022;**13**:2592. <https://doi.org/10.1038/s41467-022-30248-0>.
- Taylor JM. Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. *Stat Med* 1986;**5**:29–36. <https://doi.org/10.1002/sim.4780050106>.
- Di Tommaso P, Chatzou M, Floden EW *et al*. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;**35**:316–319. <https://doi.org/10.1038/nbt.3820>.
- Al-Chalabi A, Andrews J, Farhan S. Recent advances in the genetics of familial and sporadic ALS. *Int Rev Neurobiol* 2024;**176**:49–74. <https://doi.org/10.1016/bs.irn.2024.04.007>.
- Online Mendelian Inheritance in Man, O. (2025). Johns Hopkins University, Baltimore, MD., in press.
- van Rheenen W, van der Spek RAA, Bakker MK *et al*. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat Genet* 2021;**53**:1636–1648. <https://doi.org/10.1038/s41588-021-00973-1>.
- Gratten J, Zhao Q, Benyamin B *et al*. Whole-exome sequencing in amyotrophic lateral sclerosis suggests NEK1 is a risk gene in Chinese. *Genome Med* 2017;**9**:97. <https://doi.org/10.1186/s13073-017-0487-0>.
- Hop PJ, Kooyman M, Kenna BJ *et al*. Large-scale exome analyses reveal new rare variant contributions in amyotrophic lateral sclerosis. *Nat Genet* 2026;**58**:717–725. <https://doi.org/10.1038/s41588-026-02535-9>.
- Kratz CP. Re-envisioning genetic predisposition to childhood and adolescent cancers. *Nat Rev Cancer* 2025;**25**:109–128. <https://doi.org/10.1038/s41588-024-00775-7>.
- Witkowski L, Nichols KE, Jongmans M *et al*. Germline pathogenic SMARCA4 variants in neuroblastoma. *J Med Genet* 2023;**60**:987–992. <https://doi.org/10.1136/jmg-2022-108854>.
- Doronzo PN, Lattante S, Bernardo D *et al*. Burden of pathogenetic and likely pathogenetic variants in SPG7, SPG11 and AP4 genes in amyotrophic lateral sclerosis. *A case-control study J Neurol* 2025;**272**:455. <https://doi.org/10.1007/s00415-025-13164-3>.
- Grassano M, Koumantakis E, Palumbo FF *et al*. Genetic variants associated with neurodegenerative disorders in patients with amyotrophic lateral sclerosis and phenotypic variability. *Neurology* 2025;**105**:e214180. <https://doi.org/10.1212/WNL.0000000000214180>.
- Wu G, Chen W, Wu J *et al*. Shared burden of ultra-rare genetic variants across a spectrum of motor neuron diseases. *Transl Neurodegener* 2025;**14**:54. <https://doi.org/10.1186/s40035-025-00516-2>.
- Escherich C, Chen W, Miyamoto S *et al*. Identification of TCF3 germline variants in pediatric B-cell acute lymphoblastic leukemia. *Blood Adv* 2023;**7**:2177–2180. <https://doi.org/10.1182/bloodadvances.2022008563>.
- Price AL, Patterson NJ, Plenge RM *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38**:904–909. <https://doi.org/10.1038/ng1847>.

22. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164. <https://doi.org/10.1093/nar/gkq603>.
23. McLaren W, Gil L, Hunt SE *et al*. The Ensembl variant effect predictor. *Genome Biol* 2016;**17**:122. <https://doi.org/10.1186/s13059-016-0974-4>.
24. Baxi EG, Thompson T, Li J *et al*. Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines. *Nat Neurosci* 2022;**25**:226–237. <https://doi.org/10.1038/s41593-021-01006-0>.
25. McLeod C, Gout AM, Zhou X *et al*. St. Jude cloud: a Pediatric cancer genomic data-sharing ecosystem. *Cancer Discov* 2021;**11**:1082–1099. <https://doi.org/10.1158/2159-8290.CD-20-1230>.
26. Hudson A, Fournier M, Coulombe J *et al*. Using existing pediatric cancer data from the Gabriella miller kids first data resource program. *JNCI Cancer Spectr* 2023;**7**. <https://doi.org/10.1093/jncics/pkad079>.
27. Bonfiglio F, Lasorsa VA, Cantalupo S *et al*. Inherited rare variants in homologous recombination and neurodevelopmental genes are associated with increased risk of neuroblastoma. *EBioMedicine* 2023;**87**:104395. <https://doi.org/10.1016/j.ebiom.2022.104395>.
28. Kim J, Vaksman Z, Egolf LE *et al*. Germline pathogenic variants in neuroblastoma patients are enriched in BARD1 and predict worse survival. *J Natl Cancer Inst* 2024;**116**:149–159. <https://doi.org/10.1093/jnci/djad183>.
29. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 2009;**25**:1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>.
30. Grobner SN, Worst BC, Weischenfeldt J *et al*. The landscape of genomic alterations across childhood cancers. *Nature* 2018;**555**:321–327. <https://doi.org/10.1038/nature25480>.
31. Conomos MP, Reiner AP, Weir BS *et al*. Model-free estimation of recent genetic relatedness. *Am J Hum Genet* 2016;**98**:127–148. <https://doi.org/10.1016/j.ajhg.2015.11.022>.