The Neurodegenerative Disease Knowledge Portal: Propelling Discovery Through the Sharing of Neurodegenerative Disease Genomic Resources

Allison A. Dilliott1,2, Maria C. Costanzo3, Noël P. Burtt3, Sara Bandres-Ciga4-5, Cornelis Blauwendraat4-5, Bradford Casey6, Quy Hoang3, Hirotaka Iwaki4-7, Dongkeun Jang3, Jonggeol Jeffrey Kim4-5, Hampton L. Leonard4-7, Kristin S. Levine4-7, Mary Makarios4-5, Trang T. Nguyen3, Guy A. Rouleau1-2,8, Andrew B. Singleton4-5, Patrick Smadbeck3, J Solle6, Dan Vitale4-7, Mike A. Nalls4-5,7, Jason Flannick3,9-10, and Sali M. K. Farhan1-2,8

1Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada
2Montreal Neurological Institute-Hospital, McGill University, Montreal, QC, Canada
3Programs in Metabolism and Medical & Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA
4Center for Alzheimer’s and Related Dementias, NIH, Bethesda, MD USA
5Laboratory of Neurogenetics, NIH, Bethesda, MD, USA
6Michael J. Fox Foundation for Parkinson’s Research, NY, NY USA
7DataTecnica LLC, Washington, DC, USA
8Department of Human Genetics, McGill University, Montreal, QC, Canada
9Department of Pediatrics, Boston Children’s Hospital, Boston, MA, USA
10Department of Pediatrics, Harvard Medical School, Boston, MA, USA

Co-Corresponding Authors:
Sali M. K. Farhan
Email: sali.farhan@mcgill.ca

Jason Flannick
Email: Jason.flannick@childrens.harvard.edu

Keywords: Neurodegenerative disease, Knowledge Portal, open-science, genomic resource, genomic tools
Abstract
Although large-scale genetic association studies have proven opportunistic for the delineation of neurodegenerative disease processes, we still lack a full understanding of the pathological mechanisms of these diseases, resulting in few appropriate treatment options and diagnostic challenges. To mitigate these gaps, the Neurodegenerative Disease Knowledge Portal (NDKP) was created as an open-science initiative with the aim to aggregate, enable analysis, and display all available genomic datasets of neurodegenerative disease, while protecting the integrity and confidentiality of the underlying datasets. The portal contains 218 genomic datasets, including genotyping and sequencing studies, of individuals across ten different phenotypic groups, including neurological conditions such as Alzheimer’s disease, amyotrophic lateral sclerosis, Lewy body dementia, and Parkinson’s disease. In addition to securely hosting large genomic datasets, the NDKP provides accessible workflows and tools to effectively utilize the datasets and assist in the facilitation of customized genomic analyses. Here, we summarize the genomic datasets currently included within the portal, the bioinformatics processing of the datasets, and the variety of phenotypes captured. We also present example use-cases of the various user interfaces and integrated analytic tools to demonstrate their extensive utility in enabling the extraction of high-quality results at the source, for both genomics experts and those in other disciplines. Overall, the NDKP promotes open-science and collaboration, maximizing the potential for discovery from the large-scale datasets researchers and consortia are expending immense resources to produce and resulting in reproducible conclusions to improve diagnostic and therapeutic care for neurodegenerative disease patients.
Introduction

Neurodegenerative diseases are clinically heterogeneous and complex disorders. Given their relatively high estimates of heritability\(^1\-^4\), large-scale association studies are particularly useful for gaining a greater understanding of the pathologic mechanisms driving neurodegenerative disease processes, as demonstrated by their discoveries across neurodegenerative conditions, including in Alzheimer’s disease, amyotrophic lateral sclerosis (ALS), and Parkinson’s disease\(^5\-^12\). Yet there are still few appropriate treatment options across these diseases, and diagnoses remain a challenge, as a direct result of a lack of full understanding regarding the neuropathologic mechanisms of the diseases\(^13\-^16\). To mitigate these gaps, greater effort must be made to combine resources in the pursuit of novel discovery in neurodegenerative disease genetics.

The application of open-science and data sharing principles in the pursuit of neurodegenerative disease research motivated the deployment of the Neurodegenerative Disease Knowledge Portal (NDKP; [https://ndkp.hugeamp.org/](https://ndkp.hugeamp.org/)), consisting of 218 open-access genomic summary statistics and variant datasets of individuals across ten different phenotypic groups, including neurological conditions such as Alzheimer’s disease, ALS, Lewy body dementia, and Parkinson’s disease\(^1\). Here, we further describe the online, open-access resource including details of the available data and applications for genetic discovery and result replication in the study of neurodegenerative diseases.

Overview of the Neurodegenerative Disease Knowledge Portal

---

\(^1\) All sample sizes included herein were obtained December 2023. Annual updates to the NDKP are anticipated.
To maximize the potential for discovery from the many large, novel datasets being leveraged for various neurodegenerative disease genetic association studies, and in taking inspiration from the successes of the previously deployed Type 2 Diabetes Knowledge Portal\textsuperscript{17}, a centralized repository was assembled to securely store these datasets and make their summary results widely available to the research community. The portal was created and deployed by the developers at the Broad Institute of MIT and Harvard in collaboration with the Montreal Neurological Institute-Hospital, the Global Parkinson’s Genetics Program (GP2), and National Institutes of Health (NIH) Intramural Center for Alzheimer’s and Related Dementias (CARD) into what is now known as the NDKP.

Overall, the aim of the NDKP is to aggregate, enable analysis, and display all available genomic datasets of neurodegenerative disease, while protecting the integrity and confidentiality of the underlying datasets. The NDKP is available to the broad scientific community studying neurodegeneration seeking to unveil novel genetic associations or validate primary findings from other approaches.

**Available datasets**

Although the goal of the NDKP is to expand our genomic understanding of neurodegenerative diseases, the portal includes genomic datasets from cohorts spanning ten phenotypic groups, including: 1) cerebrovascular magnetic resonance imaging (MRI) traits, 2) COVID-19, 3) immunological, 4) metabolite, 5) musculoskeletal, 6) neurological, 7) psychiatric, 8) sleep and circadian, 9) stroke, and 10) cognitive. Across these groups, 239 sub-phenotypes are captured (Supplemental Table 1). Importantly, this allows for the cross analysis of neurodegenerative diseases with possible related and overlapping features. For example, given the known association between cerebrovascular disease features and neurodegeneration risk\textsuperscript{18–21}, the NDKP
can be used to identify genes or single nucleotide polymorphisms (SNPs) relevant in both a neurological phenotype such as Alzheimer’s disease and a cerebrovascular MRI feature such as brain microbleeds, as will be discussed further below. In total, 218 genomic datasets are currently captured within the NDKP (Figure 1), which can be subdivided into two data types: 1) genotyping studies and 2) sequencing studies (Figure 2).

*Genotyping studies*

The genotyping-based studies within the NDKP largely encompass genome-wide association studies (GWAS), which employ genotyping microarray data and are aimed at identifying associations between SNPs and traits or disease states. GWAS require large sample pools of both individuals with a phenotype of interest and non-affected controls, and depending on the type of quantitative trait, can generate an effect size per variant\(^22\). The NDKP includes datasets from 206 GWAS performed between 2010 and 2023, with the largest neurodegenerative disease dataset having been part of the 2023 GP2 GWAS that included 1,028,993 samples across multiple ancestries\(^23\) (Figure 2). Summary statistics from many of the GWAS, which include the aforementioned effect sizes per statistically significant, phenotype-associated SNP, are directly downloadable from the NDKP (https://ndkp.hugeamp.org/research.html?pageid=a2f_downloads_280).

In addition to the case-control variant calls and summary statistics, further datasets have been collated in the NDKP from the GWAS analyses. “Credible sets” include sets of variants near significant genetic association signals that are likely to include the causal variant for the signals and are generated through fine mapping of the GWAS results to further investigate genetic association signals. Additionally, “effector gene lists” encompass lists of variant-adjacent genes potentially mediating the effects of the significantly phenotype-associated variants identified
within the GWAS. Although the methods by which these genes are defined are study-dependent, their inclusion in the NDKP is important as stepping stones for experiments to further define the genes that may be truly causal for a particular phenotype.

Finally, the NDKP also has two datasets from the immunological phenotype group that encompass ImmunoChip (iChip) data. The iChip is a custom-designed Illumina Infinium microarray that includes a specific set of SNPs and small insertion-deletions that were previously associated with autoimmune and inflammatory diseases through GWAS\textsuperscript{24}. It provides the added benefit of being more cost-effective than typical exome- or genome-wide microarrays.

**Sequencing studies**

Sequencing studies include the data from whole genome, exome, or smaller panel-based next-generation sequencing (NGS) analyses. Typically, these studies include sequences from both affected (individuals with the phenotype of interest) and unaffected individuals, as the data are utilized to perform variant binned association analyses, such as rare variant burden association analysis. Generally, these approaches collapse rare variants into groups that can be dictated by a variety of factors, including, but not limited to: general genomic region, individual genes, pathways of interest, minor allele frequency, or functional consequence\textsuperscript{25}. Various methods can be used to identify associations between the rare variant groups and the phenotype of interest, such as univariate or multivariate regression models. Importantly, sequencing data not only allows for analysis of variant-level associations but also for assessment of gene- and region-level associations, using variants called from the complete sequencing of all loci within a given region of the genome.
Currently, the NDKP includes the data from ten sequencing studies, including seven whole genome and three exome datasets (Figure 2). As NGS becomes more cost-effective, we expect larger and more ethnically diverse sequencing datasets to be made available and included in Knowledge Portal initiatives.

**Neurological datasets**

The NDKP includes 33 datasets within the neurological phenotype subgroup, including 32 GWAS datasets and one exome sequencing dataset (Table 1). These neurological phenotype datasets span eight sub-phenotypes, including Alzheimer’s disease (general and late-onset), ALS, Parkinson’s disease, epilepsy, and Lewy body dementia. While 26 of the datasets are of European ancestry, one dataset has subjects of African ancestry, another has subjects of Hispanic ancestry, and five datasets represent multi-ancestry analyses. In the coming years, it is a priority for the NDKP to continue expanding the ancestral diversity of the available datasets.

**Multi-omic datasets**

In addition to the genomic datasets available through the NDKP, multi-omic data from the Common Metabolic Diseases Genome Atlas (https://cmdga.org/) are integrated through the various search pages and tools provided by the portal17, as discussed further below. In total, there are 6890 multi-omic datasets produced using a wide-variety of methods including, but not limited to, ATAC-Seq, CaptureC, ChIP-Seq, HiC, and RNA-Seq. Importantly, these datasets allow for the annotation of variants captured within the NDKP based on whether they are encompassed by regions considered accessible chromatin, binding sites, candidate regulatory elements, chromatin state, gene expression, histone modifications, or target variant predictions.

**Bioinformatics processing of the data**
Upon intake, genetic and genomic datasets are subjected to a suite of bioinformatic methods to glean additional insights from the processed and integrated results. Incoming genetic association datasets are first subjected to quality control and harmonization, including ensuring that standardized column headings are utilized, inferring missing data for non-optional columns (e.g. odds ratios can be used to infer effect sizes), and lifting over all datasets to GRCh37. We also ensure that all effect sizes are in reference to the alternate allele of GRCh37, remove variants with incompatible summary statistics for the subsequent analyses, and perform a linear regression-based effect size scaling for all quantitative phenotypes\textsuperscript{17}. Variant associations are then meta-analyzed, using the METAL algorithm that infers and accounts for sample overlap between datasets (https://genome.sph.umich.edu/wiki/METAL_Documentation#Sample_Overlap_Correction) to calculate an integrated “bottom-line” association for each variant and each trait. The bottom-line analysis can both identify novel associations that are not significant at the level of individual datasets but become significant when multiple studies are considered and identify artifactual associations that may be significant in one dataset but are not replicated in others using a fixed effect method. Using these bottom-line associations, we run the Variant Effect Predictor\textsuperscript{26} to annotate predicted variant impact and perform LD-clumping using the PLINK method\textsuperscript{27}, to group variants into genetically linked sets. We use the MAGMA algorithm per gene\textsuperscript{28}, to calculate gene-level association scores based on nearby common variant associations, and per trait, to generate lists of biological pathways whose constituent genes are enriched for genetic associations for that trait. We apply the LD score regression method (LDSC) in two calculations\textsuperscript{29, 30}. Cross-trait LDSC is used to calculate the genetic correlations between all traits, while stratified LDSC provides a measure of the enrichment of genetic association signals for
each trait within annotated genomic regions such as enhancers and promoters. Finally, we apply the Human Genetic Evidence Calculator (HuGE Calculator) across all associations to categorize the weight of evidence supporting the relevance of each gene to each trait. These methods are documented in the “Help” pages of the NDKP and were also described in detail by Costanzo et al. 17.

**Application for the study of neurodegeneration**

In addition to securely hosting large genomic datasets, one of the founding aims of the NDKP is to provide accessible workflows and tools to effectively utilize the datasets and assist in the facilitation of customized genomic analyses. To allow users to carry out these aims, the portal offers four core page types, including regional pages, gene pages, variant pages, and phenotype pages, in addition to a variety of tools to allow for more structured analyses. Across these pages and tools, users are provided with summary results derived from the genomic datasets to explore genes, genomic regions, variants, or phenotypes of interest (Figure 3).

**Exploring genes/regions of interest**

Commonly, researchers from disciplines outside the realm of computational or statistical genomics identify genes or regions of the genome of interest from experimental study, such as comparative functional genomic analyses with model organisms, protein-protein interaction analyses, or expression quantitative trait loci (eQTL) analyses. Following such studies, researchers may wish to further explore whether human genetic and genomic results support the hypotheses generated via experimentation. However, it has often been difficult for these research groups to gain access to the necessary large-scale genotyping or sequencing datasets to explore these results. Further, even if data were freely available, these researchers may have lacked the
expertise to efficiently or accurately pull the required summary-level results to substantiate their hypotheses. The NDKP has aimed to fill this gap, providing a variety of results that may be of interest both to the genetics community and to researchers outside of the discipline to explore regions of the genome and specific genes that may represent new risk loci or therapeutic targets for specific neurodegenerative phenotypes (Figure 3).

Following the search of a gene of interest, the NDKP returns a variety of gene- and variant-level association summary results. At the gene-level, the search returns both common variant and rare variant gene-wide trait associations. More specifically, these results represent phenotypes for which a genetic association exists with the genes when common or rare variants, respectively, are binned and the burden of variants within the gene are compared between a cohort of individuals with the phenotype to a control cohort that does not have the phenotype. The NDKP also returns HuGE Calculator trait scores for the gene, which are calculated from a method created by Dornbos et al. and represent the extent of human genetic evidence captured within the Knowledge Portals that supports gene-phenotype associations31. For example, the well-established ALS-associated gene, SOD1, has a calculated HuGE score of 350, representing a “compelling” level of evidence for involvement of the gene in ALS, whereas the gene LDLR, which is traditionally associated with familial hypercholesterolemia and has no known associations with ALS, has a calculated HuGE score of 1.33, representing “anecdotal” evidence.

In addition to gene-level summary results, the search of a gene in the NDKP will also return variant-level summaries, including a list of variants within the gene that have individual associations with any of the phenotypes captured across the various datasets. Similarly, the search of a region of interest will return individual variant-level trait associations that have been identified across the NDKP datasets within the given region. The search of a region of interest
will also return the Genomic Region Miner (GEM) LocusZoom tool, which visualizes variant associations with single or multiple traits. Figure 4 displays the results of searching for the region surrounding a well-known Parkinson’s disease gene, SNCA (chr 4:90,571,496-90,809,466). Using the GEM LocusZoom tool, not only can the variants with significant Parkinson’s disease associations be visualized, but if other trait-associations are suspected in the region, as is the case for Lewy body dementia in this region, additional phenotypes can be queried. A table of the variant-level data captured by the GEM LocusZoom tool is also provided by the NDKP.

In addition to the genetic association summary results computed and returned by the NDKP, a variety of gene annotations can be found from the search of a given gene or region. The search of a gene of interest returns functional associations of the gene, tissue specific gene expression data, effector gene lists, UniProt cross-reference data, and external resource links as applicable. In contrast, the search of a region of interest returns a list of genes encompassed by the region.

Exploring variants of interest

Variants of interest, such as those identified through GWAS analysis or other association studies, can also be further explored using the NDKP. The search of a variant using the dbSNP identifier will return both variant annotations and phenotype associations based on the datasets encompassed in the portal (Figure 3). More specifically, the variant search page will return information regarding the closest gene to the variant or the gene it resides within, as applicable, as well as any predicted variant consequences. The search page also returns results of a phenotype-wide association study (PheWAS), providing the statistical results that describe the level of associations of the variant with any given phenotype captured by the NDKP. For example, when a search of the NDKP is performed for the APOE e4 defining SNP, rs429358, an
unsurprising significant association is observed between the variant and an increased risk of late-onset Alzheimer’s disease (OR = 3.49, p = 4.94e-324); however, the variant is also found to be significantly associated with an increased risk of brain microbleeds (OR = 1.29, p = 7.48e-10). Importantly, individual datasets can also be specifically queried for variants of interest, which can provide interesting sources of evidence for ancestry specific analyses or variant curation exercises\textsuperscript{34}.

*Exploring phenotypes of interest*

Unlike the above examples that are often driven by experimentally derived hypotheses, a researcher may also have a phenotype or disease of particular interest for which they want to develop novel human genomic derived hypotheses. For these instances, the NDKP provides phenotype search page result summaries in addition to specific tools that allow for the exploration of genetic associations (Figure 3).

When only a single phenotype is of interest, the phenotype search page provides the greatest amount of information to the user, including both variant-level and gene-level result summaries. All datasets for which the phenotype of interest is captured are clearly outlined. At the variant level, the phenotype search page provides the top genome-wide single-variant associations, and at the gene-level, associations based on the binning of common variants within each potential gene are provided. The search also provides top gene pathway associations, cell/tissue specific genetic correlations, and effector gene predictions for the phenotype, as applicable.

Using ALS as a case study, the user will find seven datasets that capture the ALS phenotype, each of which can be further explored. As anticipated, the top single-variant association signal for ALS is an intronic SNP (rs2453555) within *C9orf72* (OR = 1.19, p = 1.78e-41), which tags a
hexanucleotide repeat expansion known to be one of the most common genetic causes of ALS\textsuperscript{35-38}. Similarly, \textit{C9orf72} (\(p = 3.04\text{e-}20\), \(\text{variants} = 24\)) is the gene with the second highest common variant gene-level association, following \textit{MOB3B} (\(p = 6.53\text{e-}29\), \(\text{variants} = 95\)), a gene located nearby \textit{C9orf72}. The NDKP also indicates that ALS is significantly associated with the acanthocytosis pathway (\(p = 3.22\text{e-}7\)) and that ALS genetic associations are significantly enriched within regions annotated as enhancers in central nervous system tissues (\(p = 7.07\text{e-}5\)).

In addition to the results captured by the phenotype search pages, the NDKP also assists with single-trait analysis through the use of the HuGE Calculator that computes HuGE scores for any given phenotype-gene combination, which are described further above (https://ndkp.hugeamp.org/hugecalculator.html?prior=0.3696).

Of additional utility, the NDKP also offers summary results and tools that allow for effective multi-trait analysis. Directly within the phenotype search page, a list of genetically correlated traits for the phenotype of interest can be found. Using Alzheimer’s disease as a case study, it is unsurprising to find a significant genetic correlation with late-onset Alzheimer's disease (\(r = 0.90\), \(p = 1.22\text{e-}70\)); however, there is also a significant genetic correlation observed with Parkinson’s disease (\(r = 0.26\), \(p = 7.80\text{e-}8\)).

In addition, the NDKP hosts three tools that allow for more detailed multi-trait analysis: 1) the Signal Sifter, 2) the Gene Sifter, and 3) the Variant Sifter. The Signal Sifter and Gene Sifter tools work in a similar manner, such that multiple traits can be queried and genetic associations relevant to two or more phenotypes will be returned, but the Signal Sifter returns regions with LD-clumped variants, while the Gene Sifter returns genes. Often the user will explore genetic correlations by beginning with multiple phenotypes that they know, or suspect, are clinically correlated.
Again, using Alzheimer’s disease as a primary phenotype, the user may wish to explore its genetic correlations with brain microbleeds based on observations of neurovascular damage that have been observed in cases of neurodegenerative disease\textsuperscript{39-41}. Indeed, upon investigation of these two phenotypes with the NDKP, the Signal Sifter returns regions of LD-clumped variants significantly associated with increased risk of both traits, including the top associated region chr 19:45,387,459-45,428,235 (Alzheimer’s associated variant rs11556505, p = 4.94e-324; brain microbleed associated variant rs769449, p = 2.52e-10) (\textbf{Figure 5A}). Similarly, the Gene Sifter returns 138 genes with significant chi-square p-values that represent a measure of overall association for the gene and both traits (\textbf{Figure 5B}). The top three of these associated genes are $TOMM40$ ($p[X^2] = 1.14e-141$), $PVRL2$ ($p[X^2] = 7.40e-137$), and $APOE$ ($p[X^2] = 2.68e-102$), expectedly, all of which are captured in the top region identified with Signal Sifter and have been previously associated with the two phenotypes independently\textsuperscript{10, 42-46}.

The Variant Sifter is the Knowledge Portal’s most recently developed tool and encompasses a wide range of capabilities, broadly allowing the user to explore variant-phenotype associations based on a range of filter options including focusing on credible sets and tissue-specific epigenomic annotations. Although there are many ways to utilize the tool, the user typically will begin with a phenotype and a region or gene of interest. The Variant Sifter returns a list of variants in the region that are associated with the phenotype that can then be filtered based on user-defined criteria. The filters allow for the identification of variants in credible sets, variants within tissue-specific regulatory region annotations of interest and variants linked to specific genes.

Recently, a novel ALS-associated gene, $KANK1$, was discovered using a rare variant association analysis approach, which not only identified an enrichment of rare variants in coding regions of
the gene in individuals with ALS but also rare variants in non-coding enhancer and promoter regions. The NDKP Variant Sifter can be used to determine if any individual variants within the surrounding regulatory regions of KANK1 demonstrate a significant association with ALS, which may be important for further functional analyses regarding this gene’s association. Following the search of the chr9:370,291-846,105 region using the Variant Sifter tool with respect to the ALS phenotype, the tool returns all variants reported within the NDKP in an association plot (Figure 6A). The variants can then be filtered based on a variety of annotations of interest to the researcher. In this case, we were interested in identifying variants within an enhancer regulatory element in the central nervous system linked to KANK1 (Figure 6B-C), and the Variant Sifter tool returned 79 variants, including one variant (chr9:504,491:A>C) demonstrating an association with ALS based on the meta-analysis of NDKP data that is approaching significance ($\beta = 0.238$, $p = 7.01e-4$; Figure 6D).

**Discussion**

As researchers and consortia continue to expend immense effort and resources in producing large-scale genomic analyses, it is essential that the greatest potential for discovery is realized from these datasets. The NDKP offers a centralized knowledge base for researchers that can act as a secure, accessible, and innovative solution for data sharing, not only achieving aims of transparency when reporting novel results, but also allowing for continued discovery in neurodegenerative disease research. Furthermore, the various user interfaces and integration of analytic tools enables consistent extraction of high-quality results at the source, for both genomics experts and those in other disciplines, while protecting the integrity of the data.

In addition to its potential for discovery, the NDKP provides a centralized hub allowing the user to replicate or further investigate findings from their own independent datasets. The generation
of “omics” datasets can be cost prohibitive, particularly considering the need for both discovery and replication subsets\textsuperscript{47}; by offering a secondary dataset for replication of novel discoveries, researchers can maximize the statistical power harnessed from their in-house data. Additionally, the breadth of data types and tools available through the portal offer the ability to further explore their findings beyond only replication of results. Our goal is to continue to add available datasets to the portal. Regular data releases are planned, at minimum on a yearly basis, including the addition of at least nine new datasets spanning multiple neurodegenerative diseases and diverse ancestral populations. We aim to prioritize the integration of additional “omic” datatypes, such as single-cell, bulk RNA-seq, and proteomics analyses, in addition to incorporating deeply phenotyped clinical datasets as they become available. Finally, we welcome new collaborations, including the opportunity to incorporate additional data, methods, and tools into the NDKP. Researchers are encouraged to contact the data intake team (amp-dcc-dat@broadinstitute.org) to discuss prospective collaborations and data deposition.

Although effort is being made to expand the portal, it is important to recognize that the NDKP currently has inherent limitations regarding the available data. Most notably, there is an overrepresentation of individuals of European ancestry, largely reflecting the lack of diversity observed across the field of genomics\textsuperscript{48}. Further, most currently available datasets represent GWAS summary statistics, which do not represent all information from the original genotyping microarray data. While offering potential for further discovery and replication of common variant signals, GWAS datasets typically do not capture rare genetic biomarkers for phenotypes of interest.

Here, we have provided a comprehensive overview of the vast amount of data included in the NDKP and their possible applications, yet there are unlimited potential use cases. Additional
detailed potential workflows have been outlined by the Knowledge Portal developers, including using the NDKP to perform multi-trait analysis, rare variant association gene-level analysis, and integrative analysis (https://ndkp.hugeamp.org/help.html?page=950). The developers have also generated ample tutorials, webinars, and presentations to aid in NDKP use (https://ndkp.hugeamp.org/help.html?page=1039). Ultimately, our goal is for available data to be accessible and easy to use for both novel discovery and replication purposes, promoting open-science and collaboration, and resulting in reproducible conclusions that will improve target discovery for neurodegenerative diseases.
Declaration of interests

H.L.L., D.V., H.I., and M.A.N.’s participation in this project was part of a competitive contract awarded to Data Tecnica International LLC by the National Institutes of Health to support open science research. M.A.N. also currently serves on the scientific advisory board for Clover Therapeutics and is a scientific founder at Neuron23 Inc, he also owns stocks.

Funding

This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging (NIA), National Institutes of Health, Department of Health and Human Services; project number ZO1 AG000534, as well as the National Institute of Neurological Disorders and Stroke. It was also supported by the Global Parkinson’s Genetics Program (GP2). GP2 is funded by the Aligning Science Across Parkinson’s (ASAP) initiative and implemented by The Michael J. Fox Foundation for Parkinson’s Research (https://gp2.org). For a complete list of GP2 members see https://gp2.org.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Publication Year</th>
<th>Cases (n)</th>
<th>Controls (n)</th>
<th>Ancestry</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer's disease GWAS</td>
<td>2019</td>
<td>71880</td>
<td>383378</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Alzheimer's disease GWAS</td>
<td>2021</td>
<td>75024</td>
<td>397844</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Alzheimer's disease GWAS</td>
<td>2022</td>
<td>85934</td>
<td>401577</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Alzheimer's disease GWAS</td>
<td>2023</td>
<td>101061</td>
<td>543127</td>
<td>Multi</td>
<td>GWAS</td>
</tr>
<tr>
<td>Alzheimer's disease family history GWAS</td>
<td>2018</td>
<td>314278</td>
<td>N/A*</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis exome case-control GWAS</td>
<td>2016</td>
<td>12577</td>
<td>23475</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis GWAS</td>
<td>2017</td>
<td>13811</td>
<td>26325</td>
<td>Multi</td>
<td>GWAS</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis GWAS</td>
<td>2018</td>
<td>20806</td>
<td>59804</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis GWAS</td>
<td>2020</td>
<td>22040</td>
<td>62654</td>
<td>Multi</td>
<td>GWAS</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis GWAS</td>
<td>2021</td>
<td>29612</td>
<td>122656</td>
<td>Multi</td>
<td>GWAS</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis GWAS</td>
<td>2021</td>
<td>27205</td>
<td>110881</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Carpal tunnel syndrome GWAS</td>
<td>2019</td>
<td>12312</td>
<td>389334</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Cognitive function GWAS</td>
<td>2016</td>
<td>3864</td>
<td>7839</td>
<td>European</td>
<td>Exome</td>
</tr>
<tr>
<td>FinnGen r8 complex disease GWAS (Alzheimer’s disease)</td>
<td>2023</td>
<td>7129</td>
<td>760059</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Global Parkinson's Genetics Program GWAS</td>
<td>2023</td>
<td>62976</td>
<td>966017</td>
<td>Multi</td>
<td>GWAS</td>
</tr>
<tr>
<td>GR@CE Alzheimer's GWAS</td>
<td>2019</td>
<td>4120</td>
<td>3289</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Handedness GWAS</td>
<td>2020</td>
<td>1470460</td>
<td>N/A</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>International League Against Epilepsy GWAS</td>
<td>2018</td>
<td>225</td>
<td>24218</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>IPDGC Parkinson's disease GWAS (females)</td>
<td>2021</td>
<td>7384</td>
<td>12389</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>IPDGC Parkinson's disease GWAS (males)</td>
<td>2021</td>
<td>12054</td>
<td>11999</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>IPDGC-UK Biobank Parkinson's disease and proxy cases GWAS (female)</td>
<td>2021</td>
<td>13420</td>
<td>90662</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>IPDGC-UK Biobank Parkinson's disease and proxy cases GWAS (male)</td>
<td>2021</td>
<td>20956</td>
<td>89660</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>IPDGC-UK Biobank Parkinson's disease GWAS (female)</td>
<td>2021</td>
<td>7947</td>
<td>90662</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>IPDGC-UK Biobank Parkinson's disease GWAS (male)</td>
<td>2021</td>
<td>13020</td>
<td>89660</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>LARGE-PD Parkinson's disease GWAS</td>
<td>2021</td>
<td>807</td>
<td>690</td>
<td>Hispanic/Latin American</td>
<td>GWAS</td>
</tr>
<tr>
<td>Late-onset Alzheimer's GWAS</td>
<td>2013</td>
<td>8572</td>
<td>11312</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Late-onset Alzheimer's GWAS</td>
<td>2019</td>
<td>21982</td>
<td>41944</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Lewy body dementia GWAS</td>
<td>2021</td>
<td>2981</td>
<td>4391</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Parkinson's disease GWAS</td>
<td>2019</td>
<td>56306</td>
<td>426424</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Parkinson's disease GWAS</td>
<td>2023</td>
<td>1200</td>
<td>2445</td>
<td>African</td>
<td>GWAS</td>
</tr>
<tr>
<td>Parkinson's disease progression GWAS</td>
<td>2019</td>
<td>4093</td>
<td>N/A</td>
<td>European</td>
<td>GWAS</td>
</tr>
<tr>
<td>Parkinson's disease progression GWAS</td>
<td>2020</td>
<td>2848</td>
<td>N/A</td>
<td>European</td>
<td>GWAS</td>
</tr>
</tbody>
</table>

Further details regarding the neurological datasets captured by the Neurodegenerative Disease Knowledge Portal can be accessed online at: [https://ndkp.hugeamp.org/datasets.html](https://ndkp.hugeamp.org/datasets.html)

*25,696 individuals had a maternal family history of Alzheimer’s disease and 14,338 individuals had a paternal family history of Alzheimer’s disease. Abbreviations: GR@CE, Genome Research at Fundacio ACE; GWAS, genome-wide association study; N/A, not applicable.
Figure 1. Genomic association datasets captured by the Neurodegenerative Disease Knowledge Portal (NDKP) from various phenotypic groups. The NDKP is comprised of genomic datasets from cohorts spanning nine phenotypic groups, including cerebrovascular magnetic resonance imaging (MRI) traits, COVID-19, immunological, metabolite, musculoskeletal, neurological, psychiatric, sleep and circadian, and stroke. Datasets include genotyping data, such as genome-wide association studies (GWAS) and ImmunoChip (iChip), and sequencing studies, such as whole genome sequencing (WGS) and exome sequencing.

Figure 2. Year of publication and sample size of the 172 genomic association datasets captured by the Neurodegenerative Disease Knowledge Portal (NDKP) from various phenotypic groups. Datasets include genotyping data, such as genome-wide association studies (GWAS) and ImmunoChip (iChip), and sequencing studies, such as whole genome sequencing (WGS) and exome sequencing. For one study that provided pre-published data with no given date of publication, the year of publication 2023 was used. Samples per dataset are presented on a log10 scale.

Figure 3. Potential uses for the data and tools encompassed within the Neurodegenerative Disease Knowledge Portal (NDKP). The NDKP aims to provide accessible workflows and tools to utilize the datasets and assist in the facilitation of customized genomic analyses. The Portal offers four core search pages and a variety of tools to provide summary results derived from the genomic datasets to explore genes, genomic regions, variants, or phenotypes of interest.

Figure 4. Genomic Region Miner (GEM) LocusZoom visualization of variant-phenotype associations within the region surrounding SNCA (chr 4:90,571,496-90,809,466). As part of the summary results returned when a region of interest is searched in the Neurodegenerative Disease Knowledge Portal (NDKP), the GEM LocusZoom tool provides a visualization of all variants within that region identified across the NDKP datasets and their associations with the regions most highly associated phenotype. In the case of the region surrounding SNCA, Parkinson’s disease represents the most highly associated phenotype. However, the GEM LocusZoom tool also allows for customized visualization of variant associations with additional phenotypes, as is shown here for Lewy body dementia.

Figure 5. Multi-trait analysis using tools integrated into the NDKP, which demonstrated genomic associations with Alzheimer’s disease and brain microbleeds. (A) Signal Sifter identified regions of LD-clumped variants significantly associated with risk of both traits. (B) Gene Sifter identified 138 genes with a significant chi-square p-value, indicating overall associations between the genes and both traits.
Figure 6. Exploration and filtration of variant associations with amyotrophic lateral sclerosis (ALS) within the surrounding region of KANK1 (chr9:370291-846105). A) Association plot of all variants identified within the meta-analyzed datasets encompassed within the Neurodegenerative Disease Knowledge Portal (NDKP) within the surrounding region of KANK1 in reference to their meta-analyzed associations with ALS. B) Filtration of variants to only those in enhancer regulatory element within the central nervous system. C) Filtration of variants to only those linked to KANK1. D) All remaining variants following the annotation filters applied using the Variant Sifter tool. The star represents a variant of interest that has been manually selected based on its association with ALS that is approaching significance.
NDKP Potential Use Cases

Gene/Region of Interest
- Gene Annotations
  - Functional associations
  - Tissue specific gene expression
  - Effector gene lists
  - UniProt cross-reference
  - External Resources
- Gene-Level Associations
  - Human Genetic Evidence (HuGE) trait scores
  - Common variant trait associations
  - Rare variant trait associations
- Variant-Level Associations
  - Genomic Region Miner (GEM) LocusZoom for single- and multi- trait associations
  - Individual variant-trait associations within gene/region

Variant of Interest
- Variant Annotations
  - Closest gene
  - Predicted variant consequences
  - Locations with tissue-specific epigenomically annotated regions
- Phenotype Associations
  - Phenotype-wide association studies (PheWAS)
  - Variant Sifter identification of phenotypic associations with annotated variants

Phenotype of interest
- Single-Trait Associations
  - Genome-wide single variant associations
  - Common variant gene-level associations
  - Human Genetic Evidence (HuGE) gene scores
  - Effector gene predictions
  - Top gene pathway associations
  - Cell/tissue specific genetic correlations
- Multi-Trait Associations
  - Global genetic trait associations
  - Signal Sifter identification of multi-trait associated regions
  - Gene Sifter identification of multi-trait associated genes
Genomic Region Miner (GEM)
The LocusZoom interactive visualization allows you to explore bottom-line variant associations for multiple phenotypes across this region. Find complete documentation here. Use the 'Add phenotypes' menu below to stack up plots for multiple phenotypes. Click the 'Filter plots and table' tab to filter variants shown in the plot(s) and table by predicted consequence, closest gene, p-value, or direction of effect.

Go to the Variant Filter to explore genetic associations, credible sets, and epigenomic annotations in this region

Parkinson's disease Variant Catalog

Parkinson's disease Variant Associations

Lewy body dementia Variant Catalog

Lewy body dementia Variant Associations
A) Signal Sifter

### Associations (Ancestry: All)

<table>
<thead>
<tr>
<th>Clump</th>
<th>Alzheimer's disease</th>
<th>Brain microbleeds, any</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs5900784</td>
<td>4.94e-324</td>
<td>1.1646</td>
</tr>
<tr>
<td>rs6733839</td>
<td>3.46e-106</td>
<td>1.0993</td>
</tr>
<tr>
<td>rs10779335</td>
<td>5.76e-42</td>
<td>0.9272</td>
</tr>
<tr>
<td>rs2508379</td>
<td>5.88e-38</td>
<td>0.9272</td>
</tr>
<tr>
<td>rs1382763</td>
<td>8.56e-31</td>
<td>0.9767</td>
</tr>
<tr>
<td>rs12151021</td>
<td>9.66e-25</td>
<td>0.9802</td>
</tr>
<tr>
<td>rs7384878</td>
<td>1.31e-24</td>
<td>0.2229</td>
</tr>
<tr>
<td>rs11761007</td>
<td>1.42e-18</td>
<td>0.2553</td>
</tr>
<tr>
<td>rs10932341</td>
<td>5.96e-17</td>
<td>0.1913</td>
</tr>
<tr>
<td>rs9381563</td>
<td>2.41e-16</td>
<td>0.9825</td>
</tr>
</tbody>
</table>

B) Gene Sifter

### Build search criteria

Filter results

Matching genes: 138
Common variant p-value: &lt;0.0000005 &lt;0.000001 &lt;0.001
Rare variant p-value: &lt;0.0000005 &lt;0.000001 &lt;0.001
HuDE Score: &lt;10 | 10-100 | 100-1000 | &gt;1000

<table>
<thead>
<tr>
<th>Gene</th>
<th>P-Val(%)</th>
<th>Trait</th>
<th>HuDE Score</th>
<th>P-Val (Comment)</th>
<th>P-Val (Trend)</th>
<th>Samples</th>
<th>Variant Sifter</th>
</tr>
</thead>
</table>