Nucleic Acids Research Advance Access published October 3, 2016

doi: 10.1093/nar/gkw893

CeNDR, the *Caenorhabditis elegans* natural diversity resource

Daniel E. Cook[1,2], Stefan Zdraljevic[1,2], Joshua P. Roberts[2] and Erik C. Andersen[2,*]

[1]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, IL 60208, USA and [2]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA

Received August 15, 2016; Revised September 20, 2016; Accepted September 26, 2016

ABSTRACT

Studies in model organisms have yielded considerable insights into the etiology of disease and our understanding of evolutionary processes. *Caenorhabditis elegans* is among the most powerful model organisms used to understand biology. However, *C. elegans* is not used as extensively as other model organisms to investigate how natural variation shapes traits, especially through the use of genome-wide association (GWA) analyses. Here, we introduce a new platform, the *C. elegans* Natural Diversity Resource (CeNDR) to enable statistical genetics and genomics studies of *C. elegans* and to connect the results to human disease. CeNDR provides the research community with wild strains, genome-wide sequence and variant data for every strain, and a GWA mapping portal for studying natural variation in *C. elegans*. Additionally, researchers outside of the *C. elegans* community can benefit from public mappings and integrated tools for comparative analyses. CeNDR uses several databases that are continually updated through the addition of new strains, sequencing data, and association mapping results. The CeNDR data are accessible through a freely available web portal located at http://www.elegansvariation.org or through an application programming interface.

INTRODUCTION

Model organisms are necessary to advance our understanding of the molecular underpinnings of biomedical traits and evolutionary processes. *Caenorhabditis elegans* is a small, free-living nematode found throughout the world. This nematode has several advantages that contribute to its power as an animal model. *C. elegans* is easily maintained in laboratory environments, has a 3 to 4 day generation time and produces ~300 offspring per generation (1). The facile genetics and large experimental toolkit have made this organism a highly productive model in addressing biological questions. Furthermore, *C. elegans* has a transparent body that enables direct observation of developmental and physiological processes (2) and strains can be frozen in liquid nitrogen indefinitely creating a long-term resource for stable genetic stocks. The species also has a small genome that is comprehensively annotated (3). These experimental advances have yielded significant accomplishments, including mapping the cellular lineage of all 959 somatic cells in the hermaphrodite (4,5), a complete wiring diagram of the nervous system (6) and crucial insights into evolutionarily conserved RNA interference (7) and cell-signaling pathways (8).

Remarkably, the majority of discoveries facilitated by the study of *C. elegans* have come from the use of a single, laboratory-adapted strain from Bristol, England known as N2 (9). Because only one genetic background has been studied extensively, we have much more to learn by using the natural diversity present within this species (10,11). To address this significant gap in our experimental toolkit, a large global population of wild strains has been collected by the *C. elegans* community and citizen scientists (12,13). These strains serve as a reservoir of natural genetic variation that can be leveraged to understand the genetic drivers of evolutionary processes and the underlying causal variation for traits relevant to biomedicine using genome-wide association (GWA) mappings. These mappings correlate genotypic variation with phenotypic differences across a population to identify quantitative trait loci (QTL) (14).

Even though a few studies have shown the utility of GWA mappings to identify the genetic variation causing phenotypic differences across the *C. elegans* species (12,13,15–17), the technique has still not been widely adopted. One explanation for the lack of GWA studies in *C. elegans* is the diverse challenges associated with several necessary steps, each of which has corresponding difficulties. First, researchers require large collections of wild strains. To ensure the fidelity of these strains, care must be taken to avoid strain confusion (18). Second, researchers must genotype this large collection of wild strains to ascertain the genotypic variation for the population. The scale of this task is cost-prohibitive and organizationally difficult. Third, the large number of independent strains must be measured for

[*]To whom correspondence should be addressed. Tel: +1 847 467 4382; Fax: +1 847 491 4461; Email: Erik.Andersen@Northwestern.edu

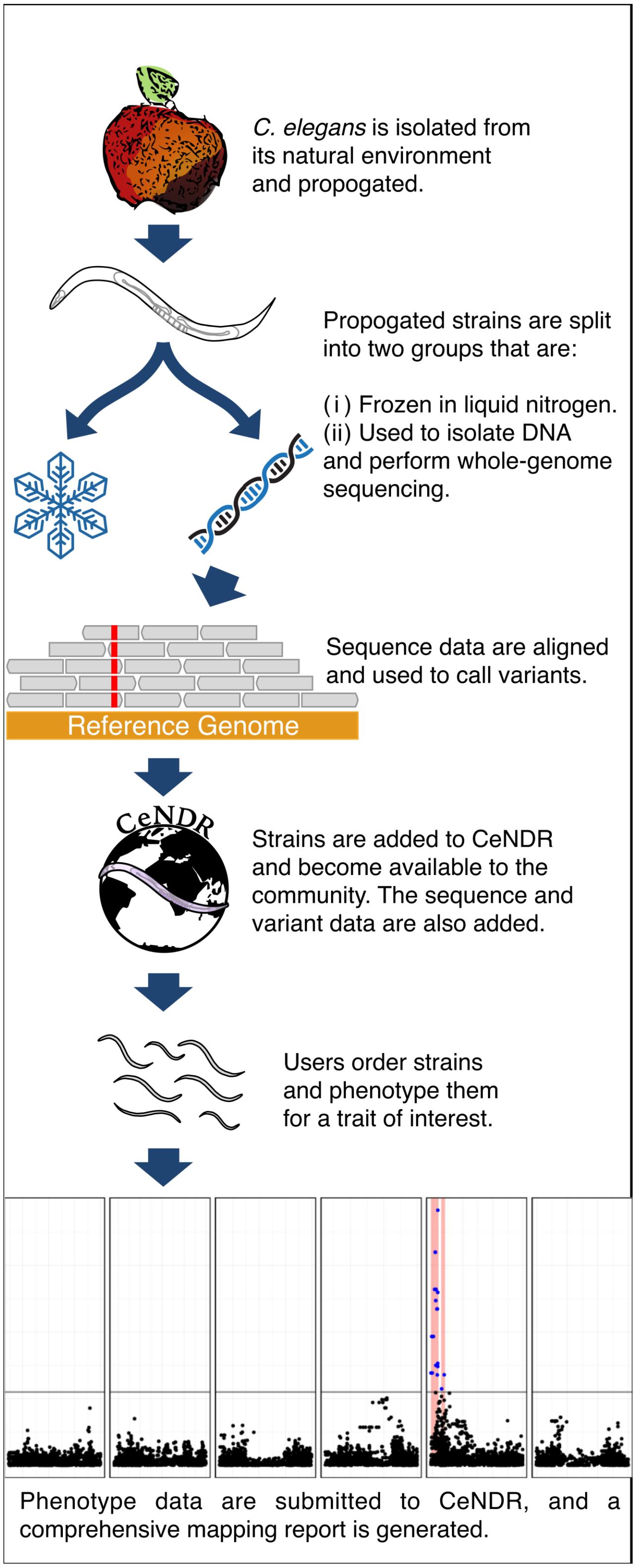
a trait of interest. Finally, researchers must correlate genotypic variation with phenotypic differences using association mapping to identify QTL. This final task requires computational skills and knowledge of statistical genetics. Altogether, these tasks require considerable laboratory, bioinformatics, and statistical expertise often performed collaboratively.

One strategy used by several model organism communities to facilitate the study of natural variation is to develop centralized repositories of strains, genotype data, and analytical pipelines that perform GWA mappings, obviating the need for laboratories to develop all of these resources independently. For example, *Drosophila* strains can be obtained from the *Drosophila* Genetic Reference Panel, a collection of genotyped inbred lines from Raleigh, NC, USA (19). In turn, these lines can be measured for a trait of interest and submitted to a web portal that performs GWA mapping (20). Similar centralized repositories and association mapping portals exist for *Arabidopsis thaliana* (21–23), and *Mus musculus* (24,25).

Here, we introduce the *C. elegans* Natural Diversity Resource (CeNDR), a comprehensive database and set of tools for examining natural variation in *C. elegans* wild strains and performing GWA mappings. CeNDR organizes metadata on natural strains, provides tools to disseminate these strains to the community, offers whole-genome sequence and variant data for each strain, and enables users to perform GWA mappings and analyze the results. CeNDR also builds upon the ideas of existing resources with an application programming interface (API). CeNDR is freely accessible without registration at http://www.elegansvariation.org. Software used to run CeNDR is open source and is available at http://www.elegansvariation.org/Software. Below, we describe how CeNDR is implemented, relevant applications, the optimized toolkit, and future plans.

IMPLEMENTATION

We have built CeNDR to facilitate the study of natural variation with three different areas of focus (Figure 1). First, CeNDR offers a platform for collecting, distributing and maintaining strains isolated from nature. Our laboratory amassed a large collection of wild strains from the *C. elegans* research community and has developed collection kits for isolating and processing additional strains. Following the receipt of new strains, a single hermaphrodite animal is propagated to ensure that the genotype is genetically distinct from a potentially heterogeneous wild population. We collect information on each strain such as its isolation location, date of collection, substrate where nematodes were found, elevation, etc. These data are integrated into the CeNDR database and can be browsed via a geographic interface on the website (Figure 2A). This dataset is also available for download or accessible through the API. After isolation and propagation of the strains, we split the population to freeze animals for long-term storage and to isolate DNA for whole-genome sequencing. This step ensures that the genotype information obtained from whole-genome sequencing can be connected directly back to a specific strain. Sample mix-ups and strain contamination (9,18) are possible when managing many strains and samples. However, our



C. elegans is isolated from its natural environment and propogated.

Propogated strains are split into two groups that are:

(i) Frozen in liquid nitrogen.
(ii) Used to isolate DNA and perform whole-genome sequencing.

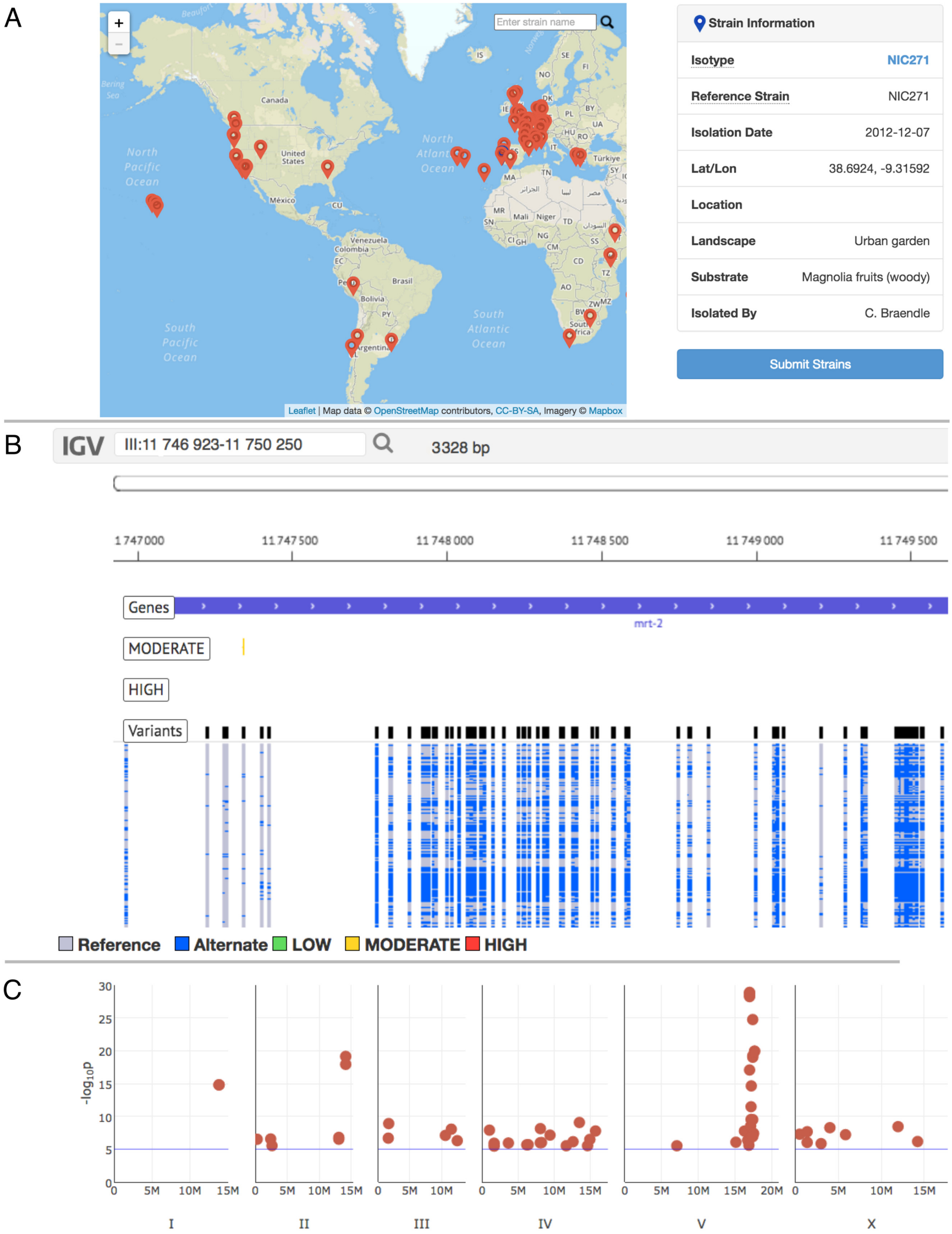
Sequence data are aligned and used to call variants.

Strains are added to CeNDR and become available to the community. The sequence and variant data are also added.

Users order strains and phenotype them for a trait of interest.

Phenotype data are submitted to CeNDR, and a comprehensive mapping report is generated.

Figure 1. Overview of the CeNDR focus areas.



Strain Information	
Isotype	NIC271
Reference Strain	NIC271
Isolation Date	2012-12-07
Lat/Lon	38.6924, -9.31592
Location	
Landscape	Urban garden
Substrate	Magnolia fruits (woody)
Isolated By	C. Braendle

III:11 746 923-11 750 250 3328 bp

Figure 2. Selected components of the CeNDR Resource. The following are screenshots of selected components of CeNDR. (**A**) A tool for interactive geographic exploration of wild isolates based on their isolation location (red markers). Additional information is displayed to the right of the map and is provided when hovering over isolation location. (**B**) A genome browser for examining genetic variation among wild isolates. Tracks for displaying genes, conservation, and the predicted effects of variants are also available. (**C**) The results from public statistically significant association mappings are added to a 'cumulative' Manhattan plot, which displays the positions of the most significant markers within a QTL confidence interval for each significant mapping.

ability to retain frozen stocks allows us to verify the genetic identity of strains should the need arise and improves the data fidelity for downstream GWA mappings.

Second, CeNDR offers whole-genome sequence and variant data of all archived wild isolates, along with metadata on gene conservation and functional studies. Most reproduction in *C. elegans* occurs through self-fertilization by hermaphrodites, resulting in the propagation of identical individuals near one another in nature. By contrast, distinct strains are sometimes found in the same isolation location. Therefore, we examine the concordance of genetic variation among strains and combine whole-genome sequence data for identical or nearly identical strains into isotypes, which represent genetically distinct genome-wide haplotypes from the same isolation location. The strain set for future GWA mapping experiments comprises a single representative strain from each isotype set. By combining sequence coverage of all strains within an isotype, we obtain high-coverage sequence data that are aligned and used to perform variant calling (see Software used for further details). All variant data are available through the API or can be downloaded in tab-delimited format or Variant Call Format files (26). Aligned sequence data is available in CRAM and BAM formats (27,28). Additionally, we have developed a genome browser for querying and visualizing genetic variation across the *C. elegans* species (Figure 2B). The genome browser allows users to toggle different tracks that detail genomic information. Available tracks include genes, conservation scores across nematode species (29) (e.g. phyloP (30) and phastCons (31)), single-nucleotide variants (SNV) identified within individual strains, and variant effects predicted with SnpEff (32).

Third, CeNDR combines whole-genome genotype data with measurements of quantitative traits to perform association mappings. The GWA mapping process is optimized for *C. elegans*, which has been used successfully in many applications (12,13,16). The GWA mapping portal is designed for non-experts and has several user-defined options along with drag-and-drop capabilities. Multiple traits can be submitted simultaneously and organized within a report, which can be kept private indefinitely, embargoed for one year, or made public. Public mapping reports that return significant QTL are added to an interactive graphic that shows all QTL identified to date (Figure 2C). CeNDR uses cloud-based virtual machines to perform GWA analyses. Results are stored in the CeNDR database, and the pipeline outputs a web-based report. Within these reports, we present users with figures, tables, interactive elements, and provide access to data in a tab-delimited format.

Additionally, we have incorporated several datasets from external sources designed to aid in comparative studies of genetic variation across diverse species and to facilitate the identification of candidate genes from GWA mappings. To query whether *C. elegans* natural variation affects genes conserved in other species, we integrated data from the Homologene database (33), associated human disease gene data listed in the Online Mendelian Inheritance in Man (OMIM) database (34), and a more nematode-focused collection of orthologs and paralogs available from WormBase (29). Once a QTL is identified, we created tools to browse the genes and potential functional connections underlying that genomic region. We integrated functional studies based on RNA interference (RNAi) screens and biochemical pathway predictions. Lastly, we developed features to enable CeNDR to interact with other services and allow access to the underlying databases through an API, which can be used to query, among other things, genetic variants, strain information, mapping report data, and *C. elegans* genes and homologs.

Software used

CeNDR website: the CeNDR website was developed using Flask (version 0.11.1). It is hosted using Google App Engine. MySQL (version 5.6.26) is used to store strain, variant, homology, and mapping data.

Sequence Analysis: raw FASTQ sequence data has been deposited under NCBI Bioproject accession PRJNA318647. Sequences were aligned to the WS245 reference genome using BWA (version 0.7.8-r455) (35). Optical/PCR duplicates were marked using PICARD (version 1.111). We used bcftools (version 1.3) to perform SNV calling (36), and SnpEff (version 4.1g) (32) to predict functional effects. Data were processed using additional scripts available at http://www.github.com/Andersenlab/vcf-kit.

Association Mapping: association mapping is performed on cloud-based virtual machines. Statistical analysis is performed using R (version 3.2.3) (37). Association mapping is performed within R using rrBLUP (version 4.4) (38). Graphics are generated using ggplot2 (version 2.0.0) (39). The CeNDR website and mapping pipelines are open source and are available on GitHub.com. See www.elegansvariation.org/software for details. We welcome community contributions.

Web-based visualization: the interactive genome browser is implemented using igv.js (version 1.0.0; github.com/igvteam/igv.js). d3.js (version 3; d3js.org), is used for certain interactive visualizations. Geographic visualizations are constructed using leaflet.js (version 0.7.7; leafletjs.com).

APPLICATIONS

Strain distribution and procurement

All wild *C. elegans* isolates in the CeNDR collection can be requested as individual strains or sets of strains. These sets are organized either into a small panel of 12 divergent strains to assess whether variation exists in a trait across the species or into several larger panels of 48 strains to measure quantitative traits for GWA mappings. Additionally, the data for each strain can be used to investigate ecological or environmental factors that influence *C. elegans*, including isolation location, substrate where the nematodes were found and the date of isolation. We also allow for anyone to submit *C. elegans* wild strains. Nematode collection kits are available from the Andersen research group and can be used to isolate new strains of *C. elegans*. As new strains are identified, they will be entered into CeNDR.

Functional studies of natural variation in *C. elegans*

Many *C. elegans* laboratories are interested in a single or small set of genes and the impacts of those genes on diverse traits. Traditional approaches used to study gene function involve the creation of loss-of-function alleles or overexpression of genes to assess phenotypic consequences. However, these methods may result in embryonic lethality or prohibit examination of more subtle aspects of gene function not observable under such extreme perturbations. For these reasons, we created tools to identify genetic variants and their predicted effects for any gene(s) of interest using a genome browser. In contrast to mutagenized strains, variants identified within wild isolates are less likely to be highly deleterious because those alleles would have been removed by natural selection if they negatively affect organismal fitness. Natural genetic variants can be integrated into a desirable genetic background, such as the laboratory-adapted strain N2 (9), using backcrossing or genome editing (40) to evaluate their effects on phenotype.

Comparative studies across *Caenorhabditis* nematodes and beyond

To investigate evolutionary processes that have occurred over longer time scales, comparative studies are often performed among different species. These studies, from *Drosophila* (41,42) to *Arabidopsis* (43), have taught us a great deal about the mechanisms of evolutionary change. Within the *Caenorhabditis* genus, comparisons of sex determination (44–48), mating behaviors (49) and gene expression regulation (50–52) are among many studies informing topics like the evolution of developmental mechanisms and behaviors. Within CeNDR, we built a homologous gene searching feature into the genome browser that can be used to identify *C. elegans* orthologs and examine genetic variation within these genes across nematodes and other species. Additionally, the genome browser includes tracks illustrating conservation using phyloP and phastCons scores across the *Caenorhabditis* genus and other nematode species. These tools allow investigators to rapidly assess whether a gene of interest has natural variation and whether that variation is in a gene region conserved across the genus. Additionally, we provide methods for researchers studying other organisms to identify homologs of their genes of interest in *C. elegans* and assess whether variation affects the functions of those genes. This tool gives non-*C. elegans* researchers an approach to test conserved gene functions in this highly tractable system.

Identifying genotype-phenotype correlations

A central goal of GWA mapping is the identification of candidate genes and genetic variants responsible for phenotypic differences across a population. We provide a GWA mapping pipeline optimized for *C. elegans* (13). This pipeline produces an easy to understand report with figures, tables, descriptions and data aimed at helping users to narrow the list of genes and variants underlying significant GWA signals. Figures include Manhattan plots (Figure 3A) that provide visualization of significance values for all markers used in the statistical test of association and plots depicting the difference in phenotype with respect to genotype at the most significant marker within a QTL confidence interval. Because *C. elegans* has linkage disequilibrium even among chromosomes (53–55), the correlation of genotype and phenotype identified on multiple chromosomes could be caused by a single region alone. Figures illustrating the linkage disequilibrium among the most significant markers from each associated region are provided to help users interpret mapping results (Figure 3B). Mapping reports also provide two interactive visualizations. The first is a map of the geographic distribution of the most significant marker with the QTL confidence interval (Figure 3C). The second interactive visualization allows users to examine Tajima's D in associated regions, which can be used to suggest whether the genotype-phenotype correlation is caused by processes under neutral, directional, or balancing selection (56). Evidence of selection at a particular locus can indicate that the QTL could have a fitness consequence in nature. Also, a list of genes within the QTL confidence interval and the predicted effects of variants within those genes are provided (Figure 3D). To integrate results obtained from the study of natural variation with the extensive knowledgebase developed from experiments using the laboratory strain, we added tools to connect identified genetic correlations to external data about gene function, including RNAi phenotypes. The genes within a QTL confidence interval can also be connected to human disease genes through the OMIM database. These diverse connections could provide additional insights into the function of a particular gene and how natural variation might affect conserved processes.

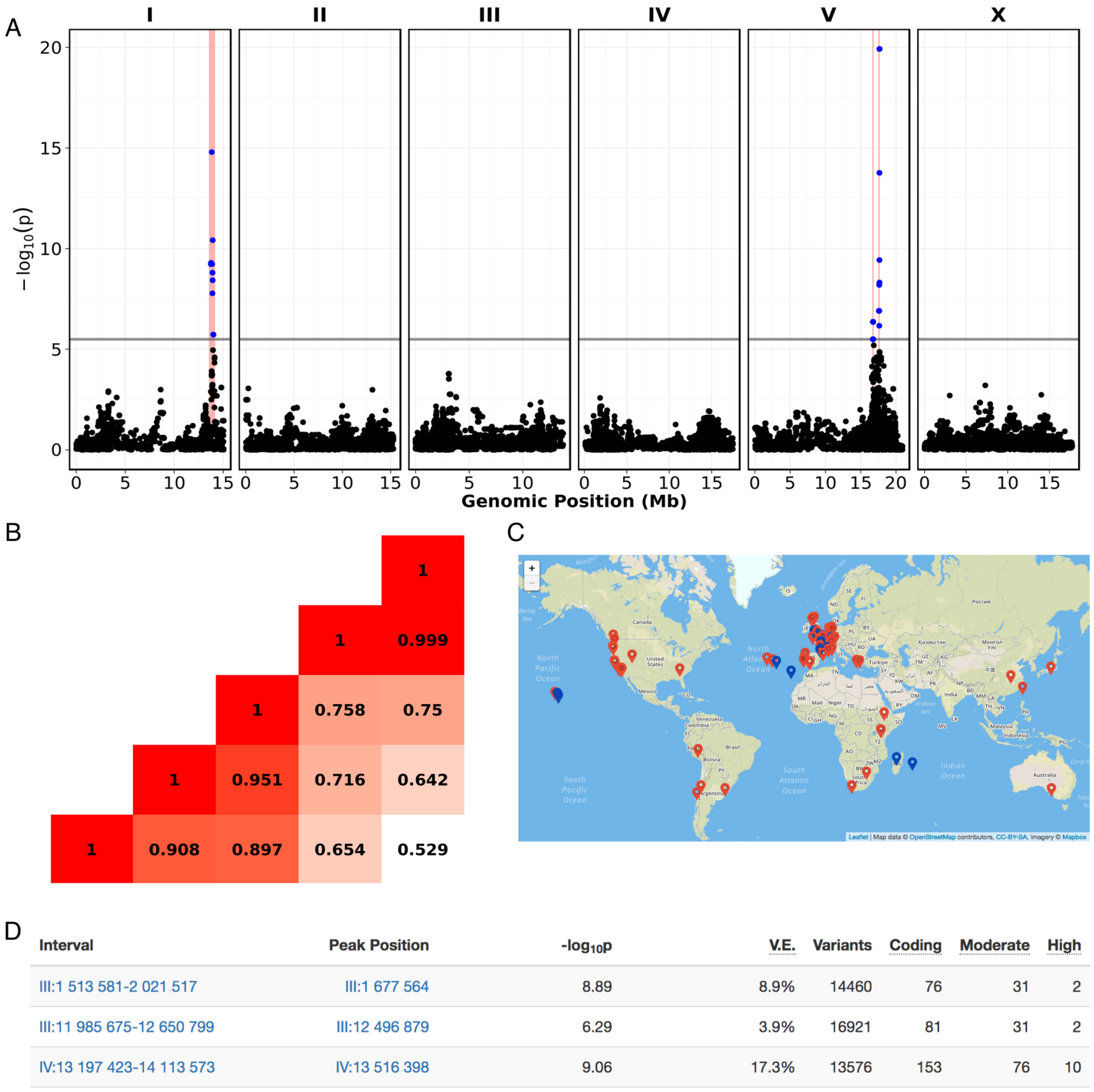
DISCUSSION

The current version of the *C. elegans* Natural Diversity Resource (CeNDR, version 1.0.0; August 2016) provides a comprehensive set of tools for examining natural variation in *C. elegans* and supports a diverse array of applications spanning studies of evolutionary processes to traits conserved with humans. History has shown that centralized resources provide numerous benefits to research communities to address important scientific questions (23,29,33,34,57). CeNDR offers reduced redundancy of data collection (e.g. whole-genome sequencing) along with consistent data collection and organization as a centralized resource. Additionally, the unification of strain management facilitates studies of natural variation across the wide *Caenorhabditis* community and beyond. Because CeNDR is built as open-source software, it benefits from additional oversight and contributions from an active research community.

CeNDR builds upon the ideas of existing platforms designed to aid studies of natural variation in several ways. First, we uniquely provide access to strains, whole-genome sequence and variant data, and a GWA mapping pipeline within a singular resource. Second, CeNDR is highly extensible by enabling access to strain, variant and mapping report data through an API. Finally, we have developed tools to apply natural variation data beyond *C. elegans*, including tools for comparative analysis of genetic variation among species.

Future directions

CeNDR will continue to grow in three important areas. First, we will incorporate more wild *C. elegans* strains, sequence their genomes and identify natural variants. Each



Interval	Peak Position	-log$_{10}$p	V.E.	Variants	Coding	Moderate	High
III:1 513 581-2 021 517	III:1 677 564	8.89	8.9%	14460	76	31	2
III:11 985 675-12 650 799	III:12 496 879	6.29	3.9%	16921	81	31	2
IV:13 197 423-14 113 573	IV:13 516 398	9.06	17.3%	13576	153	76	10

Figure 3. The GWA mapping reports within CeNDR. (**A**) Manhattan plots provide visualization of significance values for all markers used in the statistical test of association. The y-axis is the negative base 10 log of the *P*-value obtained from the statistical test of association. The x-axis is the genomic position in millions of base pairs. Markers with a -log10 *P*-value greater than the Bonferroni-corrected significance threshold (gray line) are considered to be significantly correlated with the phenotype, indicating that linked genetic variation could be causing the observed phenotypic variation. (**B**) Linkage disequilibrium among the most significant markers from each associated region is displayed. (**C**) An interactive plot of the geographic location of strains harboring either the reference or alternative marker at the most significant marker within the QTL confidence interval is shown. (**D**) A summary table of genes and other attributes within the QTL confidence interval is output. The number of protein-coding genes with variants, genes with moderate-impact variants, and genes with high-impact variants are provided.

year, we will release a new validated set of strains to increase the statistical power of GWA mappings. Second, we will integrate additional classes of natural variants beyond SNVs, including transposon insertion, insertion-deletion, copy-number and genomic rearrangement variants. These additional classes of variation will better inform predictions of functional effects and improve our mapping resolution. Third, we will release new visualization and interactive tools to mine variation, quantitative phenotypes and conservation within and beyond *Caenorhabditis*.

ACKNOWLEDGEMENTS

We would like to thank members of the Andersen laboratory for critical comments on this manuscript. We also thank the CeNDR scientific advisory panel: Marie-Anne Félix, Matthew Rockman, Ann Rougvie, and Paul Sternberg.

FUNDING

National Institutes of Health (NIH) [R01GM107227]; American Cancer Society Research Scholar Award (to E.C.A.); Amazon Web Services Research Grant (to E.C.A.); Weinberg College of Arts and Sciences starter innovation award (to E.C.A.); National Science Foundation Graduate Research Fellowship [DGE-1324585 to D.E.C.]; Northwestern University Start-up Funds (to E.C.A.). Funding for open access charge: NIH [R01GM107227].
Conflict of interest statement. None declared.

REFERENCES

1. Wood,W.B. and Others (1987) *The Nematode Caenorhabditis Elegans*, Cold Spring Harbour Laboratory, NY.
2. Corsi,A.K., Wightman,B. and Chalfie,M. (2015) A transparent window into biology: a primer on Caenorhabditis elegans. *Genetics*, **200**, 387–407.
3. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, **282**, 2012–2018.
4. Sulston,J.E., Schierenberg,E., White,J.G. and Thomson,J.N. (1983) The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.*, **100**, 64–119.
5. Kimble,J. and Hirsh,D. (1979) The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. *Dev. Biol.*, **70**, 396–417.
6. White,J.G., Southgate,E., Thomson,J.N. and Brenner,S. (1986) The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **314**, 1–340.
7. Fire,A., Xu,S., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–811.
8. Beitel,G.J., Clark,S.G. and Horvitz,H.R. (1990) Caenorhabditis elegans ras gene let-60 acts as a switch in the pathway of vulval induction. *Nature*, **348**, 503–509.
9. Sterken,M.G., Snoek,L.B., Kammenga,J.E. and Andersen,E.C. (2015) The laboratory domestication of Caenorhabditis elegans. *Trends Genet.*, **31**, 224–231.
10. Frézal,L. and Félix,M.-A. (2015) *C. elegans* outside the Petri dish. *Elife*, **4**, doi:10.7554/eLife.05849.
11. Félix,M.-A. and Braendle,C. (2010) The natural history of Caenorhabditis elegans. *Curr. Biol.*, **20**, R965–R969.
12. Andersen,E.C., Gerke,J.P., Shapiro,J.A., Crissman,J.R., Ghosh,R., Bloom,J.S., Félix,M.-A. and Kruglyak,L. (2012) Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat. Genet.*, **44**, 285–290.
13. Cook,D.C., Zdraljevic,S., Tanny,R.E., Seo,B., Riccardi,D.D., Noble,L.M., Rockman,M.V., Alkema,M.J., Braendle,C., Kammenga,J.E. *et al.* (2016) The genetic basis of natural variation in Caenorhabditis elegans telomere length. *Genetics*, **204**, 371–383.
14. Bush,W.S. and Moore,J.H. (2012) Chapter 11: genome-wide association studies. *PLoS Comput. Biol.*, **8**, e1002822.
15. Rockman,M.V. and Kruglyak,L. (2009) Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet.*, **5**, e1000419.
16. Ghosh,R., Andersen,E.C., Shapiro,J.A., Gerke,J.P. and Kruglyak,L. (2012) Natural variation in a chloride channel subunit confers avermectin resistance in *C. elegans*. *Science*, **335**, 574–578.
17. Hodgkin,J. and Doniach,T. (1997) Natural variation and copulatory plug formation in Caenorhabditis elegans. *Genetics*, **146**, 149–164.
18. McGrath,P.T., Rockman,M.V., Zimmer,M., Jang,H., Macosko,E.Z., Kruglyak,L. and Bargmann,C.I. (2009) Quantitative mapping of a digenic behavioral trait implicates globin variation in *C. elegans* sensory behaviors. *Neuron*, **61**, 692–699.
19. Mackay,T.F.C., Richards,S., Stone,E.A., Barbadilla,A., Ayroles,J.F., Zhu,D., Casillas,S., Han,Y., Magwire,M.M., Cridland,J.M. *et al.* (2012) The Drosophila melanogaster genetic reference panel. *Nature*, **482**, 173–178.
20. Huang,W., Massouras,A., Inoue,Y., Peiffer,J.,Ràmia,M., Tarone,A.M., Turlapati,L., Zichner,T., Zhu,D., Lyman,R.F. *et al.* (2014) Natural variation in genome architecture among 205 Drosophila melanogaster genetic reference panel lines. *Genome Res.*, **24**, 1193–1208.
21. Childs,L.H., Lisec,J. and Walther,D. (2012) Matapax: an online high-throughput genome-wide association study pipeline. *Plant Physiol.*, **158**, 1534–1541.
22. Seren,Ü., Vilhjálmsson,B.J., Horton,M.W., Meng,D., Forai,P., Huang,Y.S., Long,Q., Segura,V. and Nordborg,M. (2012) GWAPP: a web application for genome-wide association mapping in Arabidopsis. *Plant Cell*, **24**, 4793–4805.
23. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
24. Kang,H.M., Zaitlen,N.A., Wade,C.M., Kirby,A., Heckerman,D., Daly,M.J. and Eskin,E. (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.
25. Bennett,B.J., Farber,C.R., Orozco,L., Kang,H.M., Ghazalpour,A., Siemers,N., Neubauer,M., Neuhaus,I., Yordanova,R., Guan,B. *et al.* (2010) A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Res.*, **20**, 281–290.
26. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
27. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
28. Bonfield,J.K. and Mahoney,M.V. (2013) Compression of FASTQ and SAM format sequencing data. *PLoS One*, **8**, e59190.
29. Howe,K.L., Bolt,B.J., Cain,S., Chan,J., Chen,W.J., Davis,P., Done,J., Down,T., Gao,S., Grove,C. *et al.* (2016) WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res.*, **44**, D774–D780.
30. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
31. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
32. Cingolani,P., Platts,A., Wang,L.L.L., Coon,M., Nguyen,T., Wang,L.L.L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly* , **6**, 80–92.

33. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
34. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
35. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
36. Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.
37. R Core Team (2013) *R Core Team. R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.
38. Endelman,J.B. (2011) Ridge regression and other kernels for genomic selection with R Package rrBLUP. *Plant Genome J.*, **4**, 250–255.
39. Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*, Springer , NY.
40. Frøkjær-Jensen,C. (2013) Exciting prospects for precise engineering of Caenorhabditis elegans genomes with CRISPR/Cas9. *Genetics*, **195**, 635–642.
41. Drosophila 12 Genomes Consortium, Clark,A.G., Eisen,M.B., Smith,D.R., Bergman,C.M., Oliver,B., Markow,T.A., Kaufman,T.C., Kellis,M., Gelbart,W. *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.
42. Ranz,J.M., Castillo-Davis,C.I., Meiklejohn,C.D. and Hartl,D.L. (2003) Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science*, **300**, 1742–1745.
43. Novikova,P.Y., Hohmann,N., Nizhynska,V., Tsuchimatsu,T., Ali,J., Muir,G., Guggisberg,A., Paape,T., Schmid,K., Fedorenko,O.M. *et al.* (2016) Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.*, **48**, 1077–1082.
44. de Bono,M. and Hodgkin,J. (1996) Evolution of sex determination in caenorhabditis: unusually high divergence of tra-1 and its functional consequences. *Genetics*, **144**, 587–595.
45. Hill,R.C., de Carvalho,C.E., Salogiannis,J., Schlager,B., Pilgrim,D. and Haag,E.S. (2006) Genetic flexibility in the convergent evolution of hermaphroditism in Caenorhabditis nematodes. *Dev. Cell*, **10**, 531–538.
46. Woodruff,G.C., Eke,O., Baird,S.E., Félix,M.-A. and Haag,E.S. (2010) Insights into species divergence and the evolution of hermaphroditism from fertile interspecies hybrids of Caenorhabditis nematodes. *Genetics*, **186**, 997–1012.
47. Haag,E.S. and Kimble,J. (2000) Regulatory elements required for development of caenorhabditis elegans hermaphrodites are conserved in the tra-2 homologue of C. remanei, a male/female sister species. *Genetics*, **155**, 105–116.
48. Guo,Y., Chen,X. and Ellis,R.E. (2013) Evolutionary change within a bipotential switch shaped the sperm/oocyte decision in hermaphroditic nematodes. *PLoS Genet.*, **9**, e1003850.
49. Rene Garcia,L., LeBoeuf,B. and Koo,P. (2007) Diversity in mating behavior of hermaphroditic and male–female Caenorhabditis nematodes. *Genetics*, **175**, 1761–1771.
50. Tu,S., Wu,M.Z., Wang,J., Cutter,A.D., Weng,Z. and Claycomb,J.M. (2015) Comparative functional characterization of the CSR-1 22G-RNA pathway in Caenorhabditis nematodes. *Nucleic Acids Res.*, **43**, 208–224.
51. Barrière,A. and Ruvinsky,I. (2014) Pervasive divergence of transcriptional gene regulation in Caenorhabditis nematodes. *PLoS Genet.*, **10**, e1004435.
52. Yanai,I. and Hunter,C.P. (2009) Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved. *Genome Res.*, **19**, 2214–2220.
53. Cutter,A.D. (2006) Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer Caenorhabditis elegans. *Genetics*, **172**, 171–184.
54. Haber,M., Schüngel,M., Putz,A., Müller,S., Hasert,B. and Schulenburg,H. (2005) Evolutionary history of Caenorhabditis elegans inferred from microsatellites: evidence for spatial and temporal genetic differentiation and the occurrence of outbreeding. *Mol. Biol. Evol.*, **22**, 160–173.
55. Barrière,A. and Félix,M.-A. (2005) High local genetic diversity and low outcrossing rate in Caenorhabditis elegans natural populations. *Curr. Biol.*, **15**, 1176–1184.
56. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
57. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.