SPECIFIC AIMS

Aim 1: To investigate the time periods when exposures to environmental risk factors carry the greatest risk for later development of ALS in Northern New England and Ohio.

We will investigate ALS disease latency using our questionnaire-based databases from Northern New England (NNE) and Ohio (OH). These questionnaires include detailed records of residences where the subjects lived over their lifetime, as well as occupations, recreational activities, exposures, medical history, head injuries, military experience, diet, and smoking. We propose to continue to enroll during the three years of this new CDC grant, by the end of which we expect to have a total database of 750 ALS patients, 550 population controls, and 425 clinic controls. Additionally, we will request from the National ALS Registry demographic information and data from any completed exposure modules from subjects living in NNE and OH.

Our time-linked subject databases will be used in case-control analyses of époques of exposure to environmental toxins/toxicants carrying the greatest risk for later development of ALS. We will estimate exposure to many sources of environmental risk factors, including waterbody cyanobacteria, pesticides, Superfund and Brownfield sites, landfills and municipal incinerators. For these analyses, we have already collected spatiotemporal databases on all these sources of environmental exposures in NNE and OH, which we will use in GIS studies to estimate residential exposures to each type of pollutant in each époque for each subject. We expect these studies to reveal what periods of life carry the greatest risk for developing ALS from exposures to environmental triggers and what is the incubation period of ALS.

Aim 2: To investigate the time periods when exposures to cyanobacteria and to pesticides carry the greatest risk for later development of ALS in the United States.

The National ALS Registry currently has data on ~2700 self-enrolled patients with ALS from the whole of the USA, which includes residential information at the time of enrolling, and about an equal number that have only city and state of residence. We propose to assemble our databases of pesticide applications and of satellite remote sensing-quantified content of cyanobacteria in waterbodies across the continental USA, including Alaska and Hawaii. We will sample a nationally representative, age and gender matched group of controls from the general population, and obtain their 25-year residential history by linking to commercially available residential history data. We will use these data to conduct a case-control analysis of the relative risk of developing ALS from quantified estimates of exposures at the place of residence to cyanobacteria and pesticides.

Aim 3. To identify genetic variants conferring susceptibility to lifestyle factors and residential exposures to cyanobacteria and to pesticides as ALS risk factors.

Our goal is to identify genetic risk factors for ALS that manifest effects specifically in the presence of an environmental stressor. We already have a biobank of blood specimens collected from n=150 ALS and n=185 clinic control patients from NNE and OH, and by the end of the proposed grant period in 2021, we expect to have collected blood specimens from an additional n=175 ALS patients, n=200 clinic control patients, and saliva specimens from 200 population control subjects. In addition, the National ALS Registry currently has ~600 blood samples from self-enrolled ALS patients from across the USA, who have residential histories at the level of city and state, at minimum. We already have, or will construct during the period of this grant, databases of pesticide applications and of satellite remote sensing-quantified content of cyanobacteria in waterbodies across the continental USA, including Hawaii and Alaksa. We

will use these data and our machine learning techniques to conduct an analysis of gene x environment (GxE) interactions for the relative risk of developing ALS from quantified estimates of residential exposures to cyanobacteria and to pesticides.

Impact –We expect these studies to reveal time periods of life that carry the greatest risk for developing ALS from exposures to environmental triggers, as well as the incubation period of the disease. An interplay of environmental stressors and genetic variations, together with possible gene-synergistic effects, likely underlie the risk of developing the ALS. Identifying the environmental risk factors and periods of temporal susceptibility will lead to disease prevention, early-identification, and development of interventions to block disease progression.

RESEARCH STRATEGY

A: SIGNIFICANCE.

Introduction:

It is becoming increasingly clear that, like cancer, ALS results from a multi-step process with perhaps six molecular steps (1,2), that it is a clinic-pathological syndrome rather than a single disease (3), and that it results from the interactions of combinations of multiple genetic and environmental risk factors (4). Spatial variation in ALS incidence and mortality has been reported throughout the world, suggesting that, much like cancer, each case of ALS may result from exposure to one or several environmental factors interacting with one or more predisposing genetic factors (5,6). Currently, mutations of 30+ genes have been recognized as causing familial ALS (fALS) (2,7,8) and variants in at least 40 other genes may increase the risk of developing ALS (9-13). In addition to the non-environmental risk factors of age and gender, at least 20 environmental risk factors have been linked to ALS (14,15). Individuals who develop ALS as a result of exposure to an environmental factor are believed to do so only after years of chronic exposure. We do not know how long this exposure must continue to produce the disease or what is the period of an individual's life during which exposure carries the greatest risk. This problem not only complicates estimation of the degree of risk but also means that a relevant risk factor may have been missed because exposure occurred many years before the disease developed. There is evidence that the wash-in and wash-out periods for environmental toxins/toxicants may be as long as 20 years (16-19), and that exposures to environmental risk factors around the time of birth may be especially hazardous (20,21).

Background:

We have previously identified that living near to waterbodies that have frequent cyanobacterial blooms is a risk factor for ALS (22-27). There are two limitations of this prior work: 1) That correlation does not prove causation; and 2) That the most significant exposures probably occurred years prior to symptom onset. We can use the analogy with cancer, where for instance arsenic exposure has been reported to increase the incidence of bladder cancer only after 10 years of continuous exposure and the incidence continues to rise for 20-40 years after termination of exposure (28). There is evidence that the neurofibrillary tangles and amyloid plaques (protein aggregates) that are a hallmark of Alzheimer's disease may be formed ten or more years before death (29). This is of relevance since both ALS and Guam ALS/PDC, and a non-human primate model produced by feeding β-N-methylamino-L-alanine to vervets have protein aggregates as a significant part of their early pathology (30-32).

To investigate the time and duration of exposure that carry an increased risk of the later development of ALS we propose to study exposures to cyanobacteria and many other sources

of environmental pollutants that have occurred over >30 years prior to the development of the disease.

Over the last decade we have collected large databases of ALS patients and control subjects in NNE and OH that contain questionnaire-based information on environmental and lifestyle exposures, including details of addresses where the subjects lived going back to birth. We have collected similarly large databases of sources of environmental pollutants and cyanobacterial content of waterbodies in NNE and OH that go back over 30 years. We propose to use these databases to compare the histories of exposures to environmental risk factors n ALS patients and control subjects. Our comparison of the "exposomes" (33-36) of cases and controls will allow us to identify critical time periods (époques) of exposure that carry the greatest risk for later development of the disease, thereby investigating the disease "incubation period" or latency.

Environmental risk factors that have been implicated in ALS (37) include heavy metals [(lead (38-48), mercury (42,45,49-53)], agricultural and industrial chemicals (54-57), cleaning solvents/degreasers (58,59), aromatic solvents (60), atmospheric pollution (60,61), certain occupations (54,58,62), tobacco use (63,64), military service (65-69), head injuries (70-72), and exposure to cyanobacteria and cyanotoxins (22-24,26,27,32,73-87). The most likely exposure routes for cyanobacteria are through the food chain (78,88) and aerosolization, which may carry toxins and organisms for 10 km or more (86,89-97). Our databases of questionnaires have time-linked information on most of these risk factors, including detailed records of residences where the subjects lived over their lifetime. About 85% of these questionnaires record residential addresses back to the place of birth.

Americans move residence frequently, which adds complexity to the study of lifetime residential exposures to environmental risk factors for ALS. In 2015-2017, about 10% of the US population moved each year; however, more than 60% of those who moved stayed in the same county (98). The need to have detailed residential history when estimating residential lifetime exposures to a putative environmental pollutant is illustrated by the lifetime residential history and estimated cyanobacteria exposures of an ALS patient (see Table 1 below). The exposure is different for each residence and must be estimated from comprehensive spatiotemporal databases of environmental toxins/toxicants for each year. In fact, these changes of residence can be used to advantage in our studies when they lead to distinct periods of variable exposure.

Table 1. Example of aggregated cyanobacterial exposure calculation for a patient with ALS.

Dates	Age	Residence Latitude	Residence Longitude	Distance to Nearest Lake (km)	Annualized Cyanobacterial Exposure[a]
1953-74	Birth – 22	XX.XX	XX.XX	23.1	0
1974-80	22 – 28	XX.XX	XX.XX	4.5	1171.3
1980-84	28 – 32	XX.XX	XX.XX	12.1	25.4
1984-97	32 – 45	XX.XX	XX.XX	1.8	8175.2
1997-2008	45 – to diagnosis at 56	XX.XX	XX.XX	4.9	279.7
[a] – Kernel density estimate (KDE) of annual cyanobacterial exposure at residence. [b] – Aggregated lifetime cyanobacterial exposure in KDE units.					Lifetime[b]: 8591.6

B. INNOVATION

- The combination of our own databases of ALS patients and controls with the large dataset of cases in the National ALS Registry who lived in regions where we have granular data on residential exposures to environmental risk factors for ALS.
- Investigation of spatiotemporal-linked periods of exposures to environmental risk factors for ALS in these regions.
- Investigation of a wide range of environmental risk factors for ALS in these regions.
- Generation of environmental exposure metrics of cyanobacteria using calibrated satellite observations over space and time.
- The use of the whole database of patients from the National ALS Registry to investigate quantified estimates of residential exposure to two specific ALS risk factors, pesticides and cyanobacteria, across the whole of the United States.

C. APPROACH

Aim 1: To investigate the time periods when exposures to environmental risk factors carry the greatest risk for later development of ALS in Northern New England and Ohio.

Overview:

Our time-linked subject and environmental databases will be used in case-control investigation of époques of exposures that carry the greatest risk for later development of ALS. We will use Geographical Information Systems (GIS) software to estimate residential exposures to many environmental risk factors, including heavy metals, pesticides, Superfund and Brownfield sites, municipal incinerators, landfills and the chemicals that they release, and cyanobacteria. For these analyses, we have already collected spatiotemporal databases on all these sources of environmental exposures in NNE and OH, which we will use in GIS studies to estimate residential exposures to each type of potential risk factor in each époque for each subject. We expect these studies to reveal what periods of life carry the greatest risk for later development of ALS from exposures to environmental triggers.

Preliminary Studies:

On-going case-control studies. With funding from our current CDC grant (1 R01 TS000245-01, grant period October 1, 2016-September 30, 2018), we have already collected over >100 ALS patients and >190 random population controls, and are conducting case-control studies of residential exposures to environmental risk factors in OH. We are continuing our case-control studies of residential exposures to environmental risk factors in NNE, with ~250 ALS patients and ~150 clinic controls that were collected during our earlier ALSA grants and our CDC Contract 200-2014-59046. We continue to expand the NNE databases. We are continuing our studies of residential exposures to environmental risk factors in 1451 ALS patients in Florida collected in the Florida Surveillance Project (99). Some of the results of our studies have already been presented or published (22-27,100-103).

We have already collected large multi-matrix databases of environmental pollutants. We have already collected a total of nearly 5000 sources of environmental toxicants in NNE and OH listed below (see Table 2. Note: FL is part of another study). For each pollutant site, we have extracted lists of all chemicals released, annotated with the matrix or media (i.e. groundwater, soil, sludge**,** air etc.), contaminant level (e.g. trichlorethylene [TCE] in ppb), and date tested. We

have collected these data from the databases of the CDC's Agency for Toxic Substances and Disease Registry (ATSDR), TRI, the Geospatial Research, Analysis and Services Program (GRASP), the Superfund National Priorities List (NPL), US Environmental Protection Agency (EPA), and the State environmental and health agencies. We have also collected databases of pesticide applications in NNE and OH dating back in many instances for several decades (See Aim 2 below).

Table 2. Contents of our environmental pollutant databases.

State	Facilities or Counts[5]	Total Samples[4]	Total Individual Chemicals[1]
NH	6,192	22,570	474
VT	1,752	30,192	293
FL	2,489[3]	120,764[2]	454[2]
OH	2,551[3]	282,502	944

[1,2] Chemical sampling still to be collected for solid waste facilities[1], NPL, and Brownfield sites.
[3] Dams and wastewater treatment plants not yet collected.
[4] Samples include county pesticide application records 1992-2012.
[5] Includes US EPA TRI, US EPA NPL, US EPA Brownfields, State Agency Solid Waste, state agency dams, and state agency wastewater treatment plants.

There are numerous sites with TCE contamination in NNE and OH. As an example, Auburn Road Landfill (Londonderry, NH) had groundwater TCE levels 3 – 63 parts-per-billion (ppb) 1993-1999, and 51 ppb in 2012, and Tibbits Road Superfund NPL site (Barrington, NH) had levels in monitoring wells >5000 ppb despite remedial action (104). The EPA Maximum Contaminant Level (MCL) allowable for TCE in water is 5 ppb. We have comparable data of pollution sites and pesticide use in NNE and Ohio extending back several decades. We have separated data on each media studied, because point-source pollutants may spread a considerable distance to residences via several routes, including atmospheric transport, hydrological processes (surface or ground water), soil and sediment (105-112).

We have already collected a database of spatiotemporal cyanobacteria concentrations for all waterbodies >8 hectares in NNE and Ohio (see Experimental Design and Methods below). We will use these to construct comprehensive spatiotemporal databases of residential exposures to cyanobacteria.

Experimental Design and Methods:

Patient databases. We will have a database of ~450 ALS patients, ~200 random population control subjects and ~225 clinic controls from NNE and OH by the end of our current CDC grant in October 2018. During this new CDC grant, we shall continue the IRB-approved processes already in place in the Cleveland Clinic Foundation (CCF, Dr. Erik Pioro) and Dartmouth-Hitchcock Medical Center (DHMC, Dr. Elijah Stommel) to recruit ALS patients, population controls and clinic controls. Briefly, all ALS patients seen in CC and DHMC are invited to participate in this study, to sign informed consent forms, to allow their anonymous demographic and clinical information to be recorded and to complete our environmental questionnaire (Appendix 1). A similar process is pursued with clinic control patients in the 40 to 80-year age range with diagnoses of multiple sclerosis, brain and spinal cord tumors,

adult-onset epilepsy, and non-familial neuromuscular diseases, such as idiopathic peripheral neuropathies. Patients with neurodegenerative diseases (such as Alzheimer's and Parkinson's diseases) are excluded from the control group. At the same time, they are invited to contribute biosamples (see Specific Aim 3). Our success rate for receipt of completed questionnaires is currently 70 to 80%. Data from CC is sent to the Dartmouth Database and Epidemiology Core via the secure, password protected, web-based database Research Electronic Data Capture (REDCap) (http://project-redcap.org/). Source documents are stored at the CC. Electronic data is stored in the Dartmouth Database and Epidemiology Core.

To recruit random population control subjects, we use the United States Postal Service Delivery Sequence File (USPS DSF2), which is a comprehensive list of mail delivery addresses for 165 million households in the USA. The file includes information allowing separation of households from businesses and eliminates unoccupied houses and seasonal residences. We send the initial invitation letters from Dartmouth, and where necessary send multiple follow-up post-card reminders to invited subjects who have not responded. Recent epidemiologic studies have successfully utilized this methodology for case identification (113). In the first year of our current CDC grant (R01TS-000245), we have collected 193 completed questionnaires to date from a recent mailing of 2000 invitations.

We expect to enroll and receive completed questionnaires from a total of >175 additional ALS patients, >200 population controls and >200 clinic controls during the 3 years of this new CDC grant, for a total database for analysis of at least 750 ALS patients, 975 controls (550 population controls and 425 clinic controls).

Estimating spatiotemporal exposures to cyanobacteria as a risk factor for ALS. We have collected our own *in situ* water samples to measure cyanobacteria metrics (e.g., biovolume, phycocyanin [a pigment uniquely produced by cyanobacteria] and chlorophyll-a concentrations, colored dissolved organic material [CDOM]) and other parameters that influence cyanobacterial blooms (water temperature, pH etc.) to produce a best model of cyanobacterial concentration in target waterbodies in NNE and OH. We have also extracted relevant data from integrated government *in situ* sample databases (e.g., EPA GLENDA) and obtained historical data extending back more than three decades from regional networks, Federal and state programs. For additional calibration we also use the National Lakes Assessment (NLA) which helps refine mapping at thousands of lakes with *in situ* samples extending across multiple biomes. We use *in situ* samples to calibrate *contemporaneous* satellite remote sensing (SRS) spectral data from cyanobacteria in the same target waterbodies using a robust sampling scheme (27). We then extrapolate the calibration model to the satellite data that cover our whole study areas to create maps of cyanobacterial concentration for every waterbody with area >8 hectares (26,27,114). We produce a best model of cyanobacterial concentration at each pixel on the waterbody using empirical, analytical, shape, and fused algorithms. Harmonization of multiple satellites and advanced image processing (i.e. atmospheric correction, cloud processing, co-registration) allow for operational processing and back-casting (i.e. the generation of historical 'epochs' using NASA/USGS archives) of observations from the several satellites used for mapping (Landsat 5, 7 and 8, Sentinel-2, MERIS) (114) [We can provide details of the satellites, processing algorithms, mathematical calculations etc. for this mapping, if reviewers require them.]

We use calibrated SRS data maps to model the aerosol spread of cyanobacteria from the waterbody to the land where the subject's house was located. In our modeling, we use the estimated cyanobacterial content of waterbodies, their specific geography, and prevailing wind conditions. The spatial unit in this modeling is a *pixel* and each pixel within a waterbody (referred to as a *water pixel*) is a potential source of cyanobacterial exposure. Each pixel in the

land bordering a waterbody (referred to as a *land pixel*) represents a location that may receive aerosolized cyanobacteria and/or cyanotoxins from the waterbody. The modeling considers both concentration and distance of a land pixel from a water pixel (the closer a land pixel is to the water pixel, the more exposure it will receive), and also the number of water pixels near the given land pixel (a land pixel near 200 water pixels receives more exposure than a land pixel near 10 water pixels); and prevailing wind direction and speed, obtained from the National Climatic Data Center. The prevailing wind direction determines which pixels the aerosol is going to move into, and the wind speed determines the proportion of the aerosol that will move. Kernel density estimates (KDE) of exposure will be based on the distance decay principle of airborne dispersal from each water pixel. For this, we will use ArcHealth, a software package run as an extension of ArcGIS software (ESRI, Inc.) that was developed by Dr. Xun Shi, GIS Specialist on our present CDC study (R01TS-000245) of OH (115). Using the Density toolset in the ArcGIS Spatial Analyst toolbox, we will construct smoothed GIS maps of spatiotemporally distributed cyanobacteria exposures. This modeling produced a spatiotemporal map of the imputed cyanobacteria/cyanotoxin exposure value at each land pixel at each time point in NNE and OH. We extract from this map the total exposure in KDE units at each place of residence for each subject in each of the 30 years prior to diagnosis to create a table of annual aggregate exposure for each subject (See Table 1 above).

We propose to geocode the addresses provided in the residential history calendar covering the 30-year period prior to ALS diagnosis or an equivalent date for controls, i.e., convert them to longitude and latitude using a free Google API. We will perform an initial coarse set of analyses of a subset of ALS patients and controls who had spent the greatest part of the 30 years prior to diagnosis/enrollment living in NNE and OH. For missing data, for instance where the subject moved out of the study region to regions where we do not have cyanobacteria exposure data, a sliding time-window method will be used to impute cyanobacteria levels that are not available for a particular residential location in a given year. At each residential location, the window will move from present to years further in the past, imputing the value closest in time for missing values.

After we have completed the analysis of this subset of participants, we will return to the full list of subjects that includes those who lived outside of NNE and OH for major periods of time. When available, we will use the total US database of cyanobacteria content to produce maps of environmental toxicants from which to estimate exposure values at residences near to sources of waterbodies. We will also generate maps in Specific Aim 2 (below) to estimate exposures during times when individuals lived outside the target states of NNE and OH.

Estimating spatiotemporal exposures to environmental toxicants as risk factors for ALS. pollutants is a similar process to that described above for cyanobacterial spread via aerosolization. We have collected extensive spatiotemporal databases of environmental pollutants from publicly available websites (see Preliminary Results, Table 2 above). We will use GIS kernel density analysis to model spread of environmental toxicants from their sources. This analysis will generate a series of maps of annual exposures extending back in time as far as the databases allow (at least 30 years). We will then spatiotemporally link our cases and controls to these annual exposure maps and statistically compare the exposures of the cases with that of the controls (see Statistical Analysis and Power Calculations below).

As described above for cyanobacterial exposures, we will perform an initial coarse set of analyses of a subset of ALS patients and controls who had spent the greater part of the past 30 years prior to diagnosis/enrollment living in NNE and OH. Also, we will apply the same process to impute missing data. After that, we will return to the full database of participants, including

those who lived outside of NNE and OH for major periods of time. Where available, we will use publicly available contaminant datasets from states outside the NNE region to manually fill-in the levels of the single priority contaminant in the location and year of residence. For sources of environmental toxicants that have been closed or otherwise remediated, we will use background levels to impute values for years occurring after contamination end-dates (i.e. post-remediation).

We emphasize that we regard the analysis of *every potential environmental pollutant* to be of importance, but that for the purposes of illustration we will only describe the application of the methods to the example of TCE.

Spatiotemporal exposures to trichlorethylene (TCE) as a risk factor for ALS. First, we will map the spatiotemporal distribution of maximum point-source TCE levels in each media (soil, ground- and surface-water, air, etc.). As an example, in NH and VT our database contains abstracted reports of TCE (CAS# 79-01-6) at 73 independent sites and we have obtained measured TCE concentration in groundwater and soil.

Each point source will be represented by multiple rows of data, each containing a single maximal value for the year. The ALS cases in our study (existing and newly collected) will have diagnoses ranging from 2010 to 2021, so we will create annual maps of *maximum* concentrations of TCE starting in 1980 for NNE and OH. In later iterations, we will consider the impact of using the *median* level. Next, we will model the spread of TCE away from the point-source to estimate the contaminant exposures at each geographic location. GID kernel density estimations of exposure will be based on the distance decay principle of water-borne and airborne dispersal from the point sources of TCE. Using the Density toolset in the ArcGIS Spatial Analyst toolbox, we will construct smoothed GIS maps of spatiotemporally distributed TCE exposure estimates. Where available we will assess plume data from contaminated site reports to decide how to set the search-radius (bandwidth). Next, from the annual exposure maps for NNE and OH we will read the TCE exposure level for each residential location of each subject for each of the 30 years prior to diagnosis, as described above.

We will perform an initial coarse set of analyses restricted to participants who spent most of the 30 years prior to diagnosis/enrollment living in NNE. For missing data, for instance where the subject moved to a region where we have no environmental contaminant data, a sliding time-window method will be used to assign the contaminant levels that are not available for a specific point-source in a certain year. At each time point-source with missing data, the window will move from past to present, imputing the value closest in time into locations with missing values. We will use background levels to impute values for years occurring after contamination end-dates (i.e. post-remediation).

To begin to approach the complex problem of multiple sources of environmental toxicant and multiple potential risk-factor chemicals that they release, we will perform spatiotemporal analyses of exposure to individual categorical point sources listed above (e.g. landfills of each subtype), using the initial dataset of subjects who had lived most of the last 30 years in NNE and OH. We will then use the full dataset, deriving publicly available contaminant datasets from states outside NNE and OH to manually fill-in the levels of single priority contaminants in appropriate locations and years of residence. After we have completed the first global analysis of individual source types, we will investigate three individual chemical pollutants that have been identified as significant risk factors for ALS (lead, TCE, pesticides) and for which we have extensive databases.

Statistical Analysis and Power Calculations:

We will aggregate and compare exposure data for individual toxins (from exposure to cyanobacteria as an example), point sources of environmental toxicants and a limited library of chemicals released from these sources, in individual years before the date of diagnosis in each ALS patient, and before the date of completion of the questionnaire by each control subject. We will use the principles developed by Sabel (20,36,116-118).

We will develop a series of models of aggregate exposures as risk factors for ALS that are based on the *following series of theoretical assumptions*: 1) That the total amount of exposure over time is the relevant risk factor for ALS (primary analysis); 2) That the effect of exposure decrements with time, which we will model with a number of hypothesized rates; and 3) That there is a particular époque of greatest sensitivity to exposure as a risk factor for ALS (e.g. in the decade prior to diagnosis).

For example, applying Sabel's methodology to *theoretical assumption #1*, we will sum the estimated exposure values for each subject, for each environmental factor (source, chemical or cyanobacterial exposure) for each year. When a subject has more than one residence in a year, we will compute a pro-rated average of the estimated exposure values for that year. Since some of the subjects will have been mobile during the period of study, we consider these exposures to be independent observations. We will perform a case-control comparison of each year's aggregate exposures for each pollutant, extending back for 30 years. Then, we will sum exposure values for sets of years for ALS patients and control subjects, extending from minus 30 years to minus 1 year (e.g. for subjects enrolled in 2010, we sum residential exposure values in a series of aggregated 5- year periods from 1980 to 2009). We will assess the association between the aggregated residential exposure values and ALS risk using logistic regression analysis. Our models will use the case-control status as outcome, and aggregated estimated exposure values as predictor, adjusting for age and gender, and covariates such as smoking status and socioeconomic status as potential confounders. Questionnaire data of occupations and hobbies are available, providing ample opportunity for further covariate consideration and adjustment. With 750 ALS patients and 975 controls, we will have 80% power to detect a minimum odds ratio for ALS of 1.32 for contaminant exposure in the top quartile versus lower levels of exposure (alpha = 0.05). In our planned assessment of many independent contaminants we will adjust the alpha value for multiple comparisons (i.e. with a conservative Bonferroni correction for 200 chemicals and toxins the alpha is 0.00025 (0.05 / 200), giving us a minimum detectable odds ratio for ALS of 1.61).

For analysis of *theoretical assumption #3*, that there is a particular époque of greatest sensitivity to exposure, we will graph the odds ratios for exposure (y-axis) for in each year prior to diagnosis (x-axis). We will use this graph to identify that époque by assessing the exposure year(s) with the highest magnitude ALS association and the lowest p-value. In order to tease out the époque of greatest sensitivity, we will use the procedure of excluding exposures occurring within 5-, 10-year, 15-year etc. periods before the diagnosis or sampling date for controls, as used by Dickerson et al. (19). We will perform similar analyses of time-linked exposures to other environmental toxicants as risk factors for ALS. In further analyses we will model pairs (or triplets) of target chemicals for covariates to investigate the combined effects of several contaminants that might be synergistic as risk factors for ALS.

<u>Potential pitfalls and alternative approaches:</u>

Our published studies of the use of GIS kernel density methodology to identify environmental risk factors for ALS in NNE indicate the likelihood of success in extending these studies back in time to investigate the "incubation period" of ALS. We have already collected the spatiotemporal databases of environmental toxins/toxicants and have large databases of ALS patients and

controls in NNE and Ohio. We will continue to expand these databases of patients and control subjects during the 3-years of this new grant. We expect that our studies of exposures during époques prior to the development of ALS will greatly expand knowledge about when such exposures are most critical.
However, we acknowledge that one important drawback of our plan to construct the exposomes of cohorts of ALS patients and controls is that data may be incomplete. This is a problem for all time-linked database studies. In completing the questionnaires, ALS patients and control subjects are asked to list all the addresses where they had lived back to their place of birth. Though we have found that subjects have generally completed the list of addresses fully, some are incomplete or lack sufficient detail to ensure accurate geocoding. We attempt to fill-in these gaps by follow-up contacts with subjects offering coaching in the use of Google Earth to determine the GPS coordinates of past addresses. We will also have the opportunity to use commercially available databases of historical addresses, as proposed for Aim 2 below (e.g. using LexisNexis). Nevertheless, about 15% of our questionnaires were only asked for the five addresses prior to diagnoss. Also, some of our databases of environmental toxicants lack comprehensive information extending back more than 30 years, and none extend back in time long enough to encompass the time of birth of our participants. Therefore, our findings will likely be limited to the most recent three decades.

Another issue is that subjects may have moved to regions where we currently do not have environmental toxin/toxicant data. As stated, we propose to use publicly available contaminant datasets for states outside NNE and OH and to use our US-wide maps of cyanobacteria generated in Specific Aim 2 to attempt to manually fill-in missing data. Results from our focused analyses of NNE / OH will be used to prioritize a single chemical environmental contaminant for nationwide data collection and future analysis.

Aim 2: To investigate the time periods when exposures to cyanobacteria and to pesticides carry the greatest risk for later development of ALS in the United States.

Overview:

The National ALS Registry currently has full residential and other (more limited) environmental risk module data on ~2,700 self-enrolled patients with ALS from the continental USA, as well as Hawaii and Alaska, and data on city/state at the time of registration, age and gender for ~3,000 other ALS patients. We have collected databases of pesticide applications from publicly available sources and will construct US-wide calibrated SRS-quantified content of cyanobacteria in waterbodies across the whole USA. These US-wide databases will extend back in time for ~25 years. We will use these databases to conduct analyses of the relative risk of developing ALS from spatiotemporal estimates of exposures at the place of residence to cyanobacteria and pesticides as described in Aim 1. We expect these studies to reveal what periods of life in the last 25 years carry the greatest risk for later development of ALS from exposures to environmental triggers.

Preliminary Studies:
Excess ALS mortality was reported among workers exposed to a component of the herbicide Agent Orange 2,4-diphenoxyacetic acid (2,4-D) (RR 3.45; 95% CI 1.10–11.11) compared to others who were not so exposed but worked for the same employer (119). This motivated us to perform a national assessment of county-level ALS mortality rates in relation to county-level pesticide application levels. Our preliminary analysis of the cases in the US National Death

Register who died from ALS in the period 1999 to 2015 suggested that there was correlation between risk of developing ALS and pesticide application levels. To assess the temporal variation, we calculated the correlation with 2,4-D application levels by year for the period 1992 - 1999. The correlation is strongest for applications in the early 1990s (e.g. p-value=3.06×10^{-5}). Acknowledging the limitations of this ecologic analysis, we now propose an individual-level case-control assessment of the association between pesticide application and ALS using the residential history data on National ALS Repository participants.

Experimental Design and Methods:

National ALS patient database. We shall apply to the National ALS Registry for the use of their patient data, which currently consists of full residential histories and other (more limited) environmental risk module data on ~2,700 self-enrolled patients with ALS from the whole of the USA. More limited data on city/state at the time of registration, age and gender are available for ~3,000 additional ALS patients.

Nationally representative controls. We will sample a group of controls from the general population using the US Postal Service Computerized Delivery Sequence (CDS) File, which contains over 135 million residential addresses. Recent epidemiologic studies have utilized this methodology for control identification (113). Marketing Systems Group licenses this file, which is updated quarterly, and provides a sampling service. They use auxiliary databases to append names and demographic information, including age, gender, and race. We have used this vendor to sample 2000 general population controls for our on-going Ohio epidemiologic study CDC R01TS-000245. The file includes information that allows separation of households from businesses and eliminates unoccupied houses and seasonal residences. We will sample these 'general population control subjects' to include n=2700 individuals residing in the same broad geographic area of the entire USA as was covered by the National ALS Registry participants, matching the sampling to the location, age, gender, and racial distribution of the participating ALS cases.

Address history. The residential history of the participating National ALS Registry cases will be available to us as city and state. We propose to determine the 25-year residential history of the general population controls using a commercially available database of addresses. A limitation to this control group is that it only has reliable data back to 1985. Our approach utilizes the newfound availability of comprehensive residential histories for individuals, as well as spatiotemporal data on environmental contaminants. The large-scale databases for financial credit and marketing offer a new opportunity to ascertain residential histories efficiently.

LexisNexis (Dayton, Ohio) is a credit reporting company with an extensive address database. A systematic comparison of residential history data accessibility, completeness, and accuracy across three different vendors identified LexisNexis to be the most effective source of data for residential history, when compared with lifetime survey-reported data (120). This study assessed the vendor's match rate and determined that LexisNexis was able to find at least one address for 100% of the living participants and 80% of those who were deceased, for a 97% overall match rate. Although non-US residences are not represented, the database did include military addresses. For 85% of the lifetime addresses, the vendor database had the street name or exact point location (120).

The LexisNexis database also compared well to survey-reported address information collected from 1,099 participants in a Michigan bladder cancer study. They found that 96.8% of the LexisNexis three most recent addresses were concordant with those reported in the survey. Of

the lifetime history years, 71% were accounted for by the last three addresses reported from LexisNexis (121). Similarly, a study of 1000 National Institutes of Health-American Association of Retired Persons (NIH-AARP) Diet and Health Study participants compared LexisNexis with the prospectively collected address information from the United States Postal Service (USPS) National Change of Address product. They found a detailed address match rate of 86%, with 89% temporal accuracy for the enhanced LexisNexis service, which remained above 78% in all years tested (1995-2013) (122). Likewise, on a larger scale, the California Teachers Study cohort assessed 133,479 participants and found that 85% of LexisNexis addresses matched their questionnaire reports (123).

We will query the LexisNexis database to obtain residential address histories on the population controls sampled from the US Postal Service Computerized Delivery Sequence (CDS) File using the following identifiers: first name, last name, date-of-birth, current street address, city, state, zip code, and country. The results of the query will include all known addresses with the month and year that they represent. We will process these data to construct a dataset containing multiple rows per person, each representing a single address ('vertical' format). We will maintain the data in our HIPPA-compliant secure, web-based Research Electronic Data Capture (REDCap) database.

Sequential Residential histories. We will then process the residential addresses obtained from the commercial vendor into sequential residential histories, removing overlapping and conflicting addresses. We propose to focus on the diagnosis year minus 25 years to minus 5 years. This would be 1986-2006 for a patient diagnosed in 2011. The residential history ascertained for the proposed project will begin after 1985, the year when the most complete address information became available in the vendor's database (120).

We will assign a case-matched 'reference date' to each of the general population controls, and we will similarly compile addresses for the matching period prior to that date. We will geocode the addresses covering the 25-year period prior to ALS diagnosis or an equivalent sampling date for controls, i.e., convert them to longitude and latitude coordinates using a free Google API. We will then build the residential history, starting from the most recent address and moving backwards in time. We use the start-date of the recent address as the end-date for the prior address. We will modify these algorithms to improve performance on our participant data, as needed. Our goal is to ascertain the sequential cities and state of residence for the 25-year period prior to sampling.

Figure 1. Estimated agricultural use of the herbicide Agent Orange (2, 4-diphenoxyacetic acid) in the United States based on proprietary surveyed rates for Crop Reporting Districts, as reported by the U.S. Geological Survey Pesticide National Synthesis Project.

National databases of pesticide applications. Our nationwide database of pesticide applications will be extracted from a number of sources (Figure 1, CDC ATSDR Substances Map, NASS Agricultural Chemical Use Database, USGS Dataseries 752 and 907, National Center for Food and Agricultural Policy, and State departments of agriculture databases), in annual applications in lbs or kgs of individual pesticides at the county level or better, and with records going back at least to 1992 and in some cases going back to 1965. An exploratory ecologic analysis of pesticides as a risk factor for ALS based on ALS case rates in the National Death Register is provided in Preliminary Results section above. We now propose a more robust, individual-level, case-control analysis of the pesticide – ALS association.

National map of cyanobacteria content of waterbodies. We will construct a US-wide map of calibrated SRS-quantified content of cyanobacteria metrics (e.g., chl-a & PC) in waterbodies >8 hectares across the whole USA. Calibration will use data from the National Lakes Assessment (see Fig. 2) and the methodology described in Aim 1. The CONUS wide map will rely on 2002-2011 MERIS and 2015-2020 Sentinel-2 and use well established techniques focused on deep convolution network learning of Fluorescence Line Height, Floating Algae Index, Suspended Particulate Matter, and the Cyanobacteria Index (Figure 2). We have automated this process using “BigData” computational approaches (Pythion, C++, GDAL) and leverage super-computing at NASA Pleiades and NEX for execution. We further calibrate and validate (integrate robust uncertainty / probability) these models and outcomes using independent withheld in situ from GLENDA, NLA, EPA STORET, and partner databases (e.g., OhioEPA, U. Toledo, NOAA GLERL). We cross-calibrate and back-cast these outcomes with NASA / USGS Landsat archives (5TM, 7ETM+, 8OLI) for NNE and OH and include historical local / regional in situ and existing toxins data. These SRS databases will extend back in time for ~25 years for our application and generate epoch exposure estimates. By using time series imagery, fusing

moderate spatial resolution (Sentinel-2A & B) with high temporal frequency (MERIS, Sentinel-3) resolutions, and using robust calibration across biomes and time we will generate one of the highest quality and most comprehensive lake databases in existence. Our approach is similar to that used by Zhang et al. (124) whose results suggested that cyanobacterial exposure is a risk factor for non-alcoholic liver disease, but we will use a much richer series of satellite sources and cyanobacterial measures.

Figure 2. Example coverage of the 2012 National Lakes Assessment and their designated trophic status (n=1,175).

At the end of the project we will share with the science community all version-controlled code and coefficients on GitHub, as well as outcomes of the mapping.

Statistical analysis and power calculation:

We will create a time-series of nationwide contaminant-level maps representing the pesticide applications and estimated cyanobacterial toxin levels in each year. We will link the residential location of each individual (ALS patient or sampled control) to the temporally matched estimate of the geospatially distributed contaminant exposure. We use ArcHealth, a software package run as an extension of ArcGIS software (ESRI, Inc.). From the exposure map of the appropriate year, we derive the level of contaminant at the GPS coordinates of the residence of the participant in that year.

As our primary analysis, we will sum the levels of a contaminant for each year together to compute a compiled exposure over the entire etiologic period (diagnosis or reference date, minus 25 years to minus 5 years). When a participant has more than one residence in a year, we will compute a weighted average of the contaminant levels for that year. The participants are mobile during this period and thus we consider them independent observations. We will assess of the association between the compiled contaminant level sum and ALS risk using logistic regression analysis. Models will use case-control status of each individual as the outcome; compiled contaminant level as the predictor, with adjustment for age, gender.

With 2700 National ALS Registry ALS patients and the 2700 nationally distributed general population controls, we will have 80% power to detect a minimum odds ratio for ALS of 1.32 for contaminant exposure in the top quartile versus lower levels of exposure (alpha = 0.05). In our planned assessment of many pesticides we will adjust the alpha value for multiple comparisons (i.e. with a conservative Bonferroni correction for 200 pesticides the alpha is 0.00025 (0.05 / 200), giving us a minimum detectable odds ratio for ALS of 1.61).

Potential pitfalls and alternative approaches:

The use of the large national database of ALS patients available through the National ALS Registry is a strength of this Specific Aim 2. However, the Registry database does not represent a random sample of patients from the USA and hence is subject to the vagaries of recruitment. We have attempted to use the much more representative collection of ALS patients from the National Death Register in our Preliminary Study (see above), but unfortunately these cases do not have residential histories linked to them, so analysis was ecologic (based on address at death). We are exploring the possibility of using commercially available databases of historical addresses to link to the individual ALS cases identified via the National Death Index, an approach that could serve as an alternative means to replicate and validate the associations we observed.

The US-wide databases of pesticide applications are not as spatially granular as our databases of other sources of environmental toxicants in NNE and OH that will be used in Specific Aim 1. This may be a case of what you gain on the swings (large numbers), you lose on the roundabouts (spatial precision), but we shall not know if this is the case until we have done the analyses.

Aim 3. To identify genetic variants conferring susceptibility to lifestyle factors and residential exposures to cyanobacteria and to pesticides as ALS risk factors.

Overview:

Our goal is to identify genetic risk factors for ALS that manifest effects specifically in the presence of an environmental stressor.

We already have a biobank of blood specimens collected from n=125 ALS and n=185 clinic control patients, in NNE and OH, and by the end of this grant period in 2021 we expect to have collected blood specimens from an additional 200 ALS patients, 150 clinic control patients, and saliva specimens from 250 population control subjects. In addition, the National ALS Registry currently has ~600 blood samples from self-enrolled ALS patients from across the USA who have residential histories at the level of city and state or more detailed. We already have, or will construct during the period of this grant, databases of pesticide applications and of satellite remote sensing-quantified content of cyanobacteria in waterbodies across the whole USA. We will use these data to conduct an analysis of the relative risk of developing ALS from quantified estimates of residential exposures to cyanobacteria and to pesticides. We note that, unlike Aims 1 and 2, in Aim 3 we will not attempt to investigate genetic interactions with environmental exposures going back in time before diagnosis because this would add an additional layer of computational analysis that is not justified at this stage.

Experimental Design and Methods:

Biosample collection. Consenting ALS patients and clinic controls in our on-going epidemiologic studies in NNE and OH (see Specific Aim 1) are invited to contribute a blood sample for which they will be offered an extra incentive payment of $20. The research sample of venous blood is collected in royal blue EDTA blood collection tubes from ALS and clinic control patients by hospital laboratory staff at the time of a blood-draw being performed for clinical purposes. Samples are maintained at 4°C and are processed within 24 hours by centrifugation to separate the red blood cells, lymphocyte, and plasma layers, which we aliquot and freeze at -80 °C. We are now requesting funding to isolate DNA from our biobanked samples from ALS patients and controls, and from additional patients and controls that we will collect during the period of this grant. We have DNA on n=150 ALS patients, n=185 controls, and we now propose enrollment of a minimum of an additional n=175 cases and n=400 controls during the study period, for a total of n=325 ALS patients, n=585 control DNA samples.

DNA isolation from blood will be performed on the lymphocyte layer using Qiagen genomic DNA extraction kits. We will obtain saliva samples from the general population controls by mailing an Oragene DNA self-collection kit to the home of study participants who have returned a questionnaire. Participants will indicate their interest in contributing a saliva sample on the consent form and will be offered an extra incentive payment of $20, payable upon our receipt of the saliva sample. The Oragene DNA kits are designed for mailing samples in population-based epidemiologic studies and contain a stabilizing reagent that preserves the DNA for several

months regardless of storage temperature. We will isolate the DNA using PrepIt-L2P kits optimized for saliva processing from DNA Genotek.

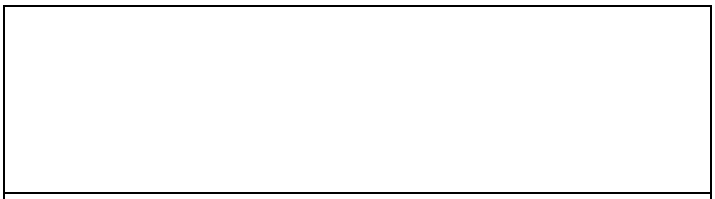

Figure 3. NeuroChip panel variants associated with ALS/FTD (adapted from Blauwendraat et al. 2017).

Genotyping. Dr. Traynor and colleagues have published genome-wide analysis data on 12,663 ALS patients and 53,439 controls (8), which will be available for comparison with our results. In addition, he has already run n=600 National ALS Biorepository DNA samples on the NeuroChip (Figure 3) (125) and assessed their C9ORF status by repeat prime polymerase chain reaction. The NeuroChip is an Illumina genotyping chip that assays 306,670 SNPs across the whole genome as a backbone, plus 179,467 variants implicated in neurological diseases. Among the variants, there are 601 that have been specifically annotated for association with ALS (Figure 3). We now propose to purchase NeuroChips and reagents and send the archived biosamples to Dr. Traynor who will perform the same genotyping assays on our NNE/OH DNA samples (see letter of commitment from Dr. Traynor).

As preliminary work, Dr. Traynor's lab has already assessed C9ORF72 on a sequentially selected set of DNA samples from n=20 of our ALS patients. Genotype data will be clustered using Genome Studio and then cleaned prior to imputation following standard practices (remove samples with low call rate, high heterogeneity, gender mismatch, or different ancestry; remove SNPs that are monomorphic, palindromic, low call rate, low minor allele frequency, or displaying evidence of Hardy-Weinberg disequilibrium SNPs).

We will take the cleaned SNP data generated on NeuroChip and impute it with up to ~12 million SNPs using the Haplotype Reference Consortium (HRC version r1.1) as a reference dataset as implemented on the Michigan Imputation server (https://imputationserver.sph.umich.edu/). Briefly, this pipeline employs Eagle2 to perform phasing, followed by imputation using miniMac software. This approach reliably captures all variation in the human genome down to a minor allele frequency of 1%.

Environmental exposure assessment. We propose a pooled analysis of exposure variables, including smoking status, occupation, and exposure to cyanobacteria and pesticides. We will identify the genetic variants interacting with each of these individual exposures that increase ALS risk. The exposure status of each participant will be ascertained using questionnaire and residential data, as described in Aims 1 & 2. We will carefully choose exposures that are reliably ascertained across the three datasets: the NNE/OH epidemiologic study questionnaires of ALS patients, clinic controls and population controls, and the National ALS Biorepository ALS

patients with completed demographic and environmental exposure modules (see Aim 2 for details).

In addition to the ALS cases and control samples that will be collected for this project, we also have access to existing genomic datasets that can be incorporated to increase power of our analysis. For example, we have access to whole-genome sequencing data on 1,200 nationally representative controls who were analyzed as part of recent Mega-GWAS ALS-1 project led by Dr. Traynor (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000101.v5.p1). These samples are obtained from individuals who undergo in-depth phenotyping and who are followed longitudinally. These samples were selected for genome sequencing as they have no personal medical history or first degree relative with Alzheimer's or Parkinson's disease, ALS, ataxia, autism, bipolar disorder, brain aneurysm, dementia, or dystonia. Because of our interest in geospatially related environmental factors, we will also exclude those participants who have a spouse with neurological illness. Participants in this study reported age, gender, ethnic and racial categories, as well as current zip code. Smoking history was collected as never, previous, current, and years of smoking.

Specifically, we will harmonize the data on smoking across the three datasets to ascertain never, former, or current smoking status. Zip code and the city and state of residence will be used to estimate exposure to cyanobacteria and pesticides using the US-wide databases collected in Aim 2.

The final pooled genotype / environmental dataset will incorporate the data from our studies: ALS patients (n=325), clinic and population control subjects (n=585), as well as existing whole genome genotyping data on n=600 National ALS Biorepository DNAs and n=1200 additional genotyped controls. The anticipated pooled sample size for GxE analysis is thus n=925 ALS cases and n=1785 controls.

Analysis of genetic variant x environment (GxE) interactions. We will employ the machine-learning method Multifactor Dimensionality Reduction (MDR) to identify GxE interactions, a technique we have used before with sparse datasets (126-129). The MDR software is open-source and freely available from http://www.epistasis.org. The method estimates the ratio of cases to controls for each 2-way combination of factors making a high-risk or low-risk determination in each of the 9 cells based on the ratio cases vs. controls. For example, wild-type, heterozygous, variant genotypes by never, former, or current smoking status. We will also investigate gene-gene interactions (epistasis) and their possible synergistic association with ALS disease risk for environmental contaminant exposures in the setting of genetic predisposition.

MDR defines a single variable that incorporates information from several genetic loci and/or environmental factors that can be divided into high risk and low risk combinations. This new summary variable is evaluated for its ability to classify and predict outcome risk status using cross-validation and permutation testing. We will select the best MDR model as the one with the lowest average prediction error. An error rate of 50% is expected under the null hypothesis. Statistical significance is determined using permutation testing. Here, the case-control labels are randomized 1000 times and the entire MDR model fitting procedure repeated on each randomized dataset to determine the expected distribution of testing accuracies under the null hypothesis. It is the combination of cross-validation (e.g. 10-fold) and permutation testing (1000-fold) that reduces the chances of making a type I error due to multiple testing (130,131).

To avoid over-fitting, analyses will be iteratively performed across multiple cross-sections of the study population. The statistical significance is calculated by Monte Carlo permutation using the cross-validation consistency.

We will then evaluate the multiplicative vs. additive nature of the interactions between the SNP allele and environmental factor combinations predicted by MDR using the likelihood ratio test by including interaction terms in a logistic regression model, with adjustment for age, gender and other potentially confounding covariates. Statistical significances of the interactions are assessed using likelihood ratio tests comparing the models with and without the interaction terms (e.g. logit(p) = constant + SNP1 + exposure vs logit(p) = constant + SNP1 + exposure + interaction). MDR identifies combinations of predictive factors but does not distinguish between multiplicative and additive effects.

Functional grouping of genes with SNPs. The Gene Set Analysis (GSA) method will be performed to assess the significance of the each of the functional groups of genes in relation to ALS risk. GSA ranks the SNPs by their correlation with the endpoint (positive for decreased risk, negative for increased risk). An enrichment score is calculated as a running-sum statistic that increases when the next SNP down the list is in the same functional group but decreases when it is not. Thus, enrichment scores with the maximum deviation from zero are achieved when multiple SNPs of the same functional group are ranked close together at the very top or bottom of the ranked correlation list. The peak enrichment score is the highest positive score, indicating that the variant form of the SNP is protective, or the lowest negative score, indicating that the variant form of the SNP is a risk factor. The False Discovery Rate (FDR) is then calculated using 1000 permutations of the endpoint to estimate the probability that the enrichment score represents a false positive finding. The enrichment score for GSA is a "maxmean" statistic, computed by averaging the positive parts of each Z-score in a given pathway, as well as the negative parts. We choose the Z-score that is larger in absolute value.

Potential pitfalls and alternative approaches:

We do not anticipate problems with the generation of the genotyping data. Genotyping will be performed by our experienced collaborator, Dr. Traynor, based at the National Institute on Aging, who performed the genotyping on the National ALS Registry samples using the same platforms. We do not anticipate problems with sample preparation, the quality of genotype data, or the turn-around time for completion of the genetic experiments. Dr. Traynor will re-genotype samples that do not achieve the necessary coverage at no additional cost.

The type of machine learning analysis proposed is state of the art and is challenging to execute correctly. However, we are highly experienced in this type of analysis (126-129)and we are confident that it can be completed reliably and on schedule.

Given the exploratory nature of the study, we will not provide *a priori* statistical power calculations. Instead a convenience sample of 925 ALS patients and 1785 matched control subjects was chosen, and we will calculate *a posteriori* our statistical power for each exposure and genetic factor. Nevertheless, this cohort (with its clinical and environmental data, including genotype data, smoking status and residential location-based cyanobacterial and pesticide exposure estimates) is one of the largest available anywhere in the world. Our inability to provide power analysis reflects the cutting-edge nature of this project. To the best of our knowledge, no one has ever attempted a study of this scope and magnitude before, so there are no data on which to base the power calculations. In the future, other researchers will have the

possibility to add their cohorts of cases and to confirm our findings or generate novel observations. In this way, the power of our resource will grow over time.

Analysis of large genomic datasets may produce false-positive results or there may be artefacts in the data that lead to incorrect conclusions. To minimize this effect, we expect to collect several hundred additional samples to expand our database. Also we will attempt to confirm genetic and environmental factors that we find to interact by comparison with an independent cohort of Italian ALS cases and controls, available through Dr. Traynor.