HHS Public Access

Author manuscript

Brain Imaging Behav. Author manuscript; available in PMC 2015 May 28.

Published in final edited form as:

Brain Imaging Behav. 2015 March ; 9(1): 89–103. doi:10.1007/s11682-015-9354-z.

Interacting with the National Database for Autism Research (NDAR) via the LONI Pipeline Workflow Environment

Carinna M. Torgerson[1], **Catherine Quinn**[1], **Ivo Dinov**[2], **Zhizhong Liu**[1], **Petros Petrosyan**[1], **Kevin Pelphrey**[3], **Christian Haselgrove**[4], **David N. Kennedy**[4], **Arthur W. Toga**[1], and **John Darrell Van Horn**[1]

John Darrell Van Horn: jvanhorn@usc.edu

[1]Laboratory of Neuro Imaging and The Institute for Neuroimaging and Informatics, Keck School of Medicine of USC, University of Southern California, 2001 North Soto Street – SSB1-Room 102, Los Angeles, CA 90032, Phone: (323) 442-7246

[2]University of Michigan School of Nursing, 400 North Ingalls Building, Ann Arbor, MI 48109-5482

[3]Yale School of Medicine, Yale Child Study Center, PO Box 207900, 230 South Frontage Road, New Haven, CT 06520-7900

[4]University of Massachusetts Medical School, Department of Psychiatry, 55 Lake Ave., North, Worcester MA 01605

Abstract

The National Database for Autism Research (NDAR) seeks to gather, curate, and make openly available neuroimaging data from NIH-funded studies of autism spectrum disorder (ASD). NDAR has recently made its database accessible through the LONI Pipeline processing environment to enable large-scale analyses of cortical architecture and function via local, cluster, or “cloud”-based computing resources. This presents a unique opportunity to overcome many of the customary limitations to fostering biomedical neuroimaging as a science of discovery. Providing open access to primary neuroimaging data, workflow methods, and high-performance computing will increase uniformity in data collection protocols, encourage greater reliability of published data, results replication, and broaden the range of researchers now able to perform larger studies than ever before. To illustrate the use of NDAR and LONI Pipeline for performing several commonly performed neuroimaging processing steps and analyses, this paper presents example workflows useful for ASD neuroimaging researchers seeking to begin using this valuable combination of online data and computational resources.

Introduction

As with many disorders linked to the human brain, autism spectrum disorders (ASDs) seem to arise from a multitude of contributing factors – from genes to neurotransmitters and structural abnormalities (Atkinson and Braddick 2011, McPartland, Coffman et al. 2011). It is this heterogeneity that has made understanding this assortment of linked disorders

The authors declare no actual or potential competing conflicts of interest.

particularly challenging. It has also led to varying treatments, the success of which appears inconsistent among different types of autism patients. The National Autism Center (NAC; http://www.mayinstitute.org/news/press_releases.html?year=2013&id=1394; (National Autism Center 2011)) reports that many varieties of biomedical and neuropsychological data are needed to properly assess an individual patient's severity, as well as the contributing factors to the disorder before an effective treatment plan can be created. Such data types include, but are not limited to: genomic, biographical, co-morbid conditions, neuroimaging, phenotypic, gestational age, onset and latency of the disorder, onset and type of treatment, and many others.

Since 2007, the NIH has been committed to coordinating ASD research across the country by fostering collaboration between research centers in order to drive a coordinated effort towards greater understanding of the underlying biological mechanisms involved. As a result, the NIH has established a centralized national database known as the National Database for Autism Research (NDAR; http://ndar.nih.gov). NDAR is a secure research data repository specifically designed to promote scientific data sharing and collaboration among autism spectrum disorder investigators (Hall, Huerta et al. 2012). The database seeks to increase efficiency in the research of ASDs through data sharing and harmonization.

To encourage neuroimaging and genetics data deposition into NDAR, all NIH-funded investigators are required to share their brain imaging data as a condition of their research grant support. Additionally, the NIH has established the Autism Centers of Excellence (ACE) Program, which presently includes six research centers and five research networks, funded through P50, R01, and U01 grant mechanisms (http://www.nichd.nih.gov/research/supported/Pages/ace.aspx). In each network, participating ACE institutions are expected to provide all of their raw collected data to a data coordinating center (DCC) affiliated with each program, which then 1) synthesizes, checks for quality and artifacts, and 2) ensures successful upload of that data to both local as well as the NDAR national repository. Currently, NDAR compiles phenotypic, genomic, neuroimaging, neuro-signal recordings and demographic data, and as of July 2014, the database contains data from over 77,000 individuals. While a significant portion of this number represents clinical and gene sequencing data, this also includes 4,745 subjects some of whom have multimodal neuroimaging data. Data come largely from the NIH Autism Centers of Excellent (ACE) program with additional phenomic data coming from Simons Foundation Autism Research Initiative (SFARI; http://sfari.org), and the Interactive Autism Network (IAN; https://www.ianresearch.org).

One unique facet of NDAR's database is its use of a global unique identifier (GUID). This GUID is obtained by entering relevant personal identifiable information (PII) from a subject's birth certificate into NDAR's GUID Tool software, or by sending the PII to a data coordinating center for GUID generation. This information will assign the subject a unique GUID that stays with that subject's NDAR record throughout their life. If the subject participates in a longitudinal study, or enrolls in a study at a different location at a later date, their GUID will be re-generated as the same number, so that all their data can be systematically aggregated. This GUID, however, cannot be used in reverse to obtain personal identifiable information, thereby preserving the anonymity and confidentiality of

the subject in line with HIPAA guidelines. Access to anonymized data may then be provided for free to qualified investigators - who are registered with the ERA commons (https://commons.era.nih.gov) and also secure local institutional approval (see http://ndar.nih.gov/ndarpublicweb/access.html for details on the access request process). Once approved by NIH officials (http://ndar.nih.gov/policies_data_access_committee.html), researchers are able to search NDAR contents, create data packages of useful de-identified data sets, and download them for further examination.

Database compilation of primary data in a manner such as this appears to be an emerging trend in the future of NIH-funded neuroimaging research. A 2004 paper by the directors of six neuroscience institutes at the NIH posits that three areas in neuroscience research in particular could be greatly improved and accelerated by large-scale database sharing: genetics, imaging technology, and clinical research (Insel, Volkow et al. 2004). The inability to superimpose data from one modality to another, they claim, prevents the analysis of important existing relationships in brain function. When tackling problems that require large amounts of information, such as co-morbid disorders or genetics, one resource alone will not be sufficient to capture the whole picture of what factors are at play. Research funding is much more efficiently used if its effects are not limited to the individual research center that is awarded the grant. Or, as Insel et al. (2004) note, “large scale science is only worth the investment if it enables progress from a broad community of scientists.” Creation of freely available databases containing primary data spreads the cost of neuroscience research across a broader portion of the research community. Modern concepts for sound research require that the research be reproducible by anyone who uses the same data and follows the same methods of the original study. However, in the world of competitive, grant-funded research, financial and temporal constraints make it infeasible to actually reproduce let alone replicate another author’s results. Data-sharing removes the burden of collecting analogous data and recreating the method of analysis (Breeze, Poline et al. 2012, Poline, Breeze et al. 2012).

With that in mind, making data openly available is merely the first step toward advancing greater data reuse, re-purposing, and the performing of neuroimaging “mega-analyses” by members of the broader community. What are also needed are means to feed the data available from resources such as NDAR directly into software designed for highly efficient processing and analysis. Moreover, since many potential investigators lack access to the scale of computing resources that would allow for efficient data analyses, those analyses should be directed to remote computing environments ideally suited for large-scale processing, scalable from individual subjects, to hundreds, and eventually thousands of individuals. Combinations of such resources with direct access to primary data represent a critical component in the future of neuroscience research (Van Horn, Dobson et al. 2006) and will help advance and improve the way in which that research is conducted and reported (Breeze, Poline et al. 2012).

In what follows, we seek to illustrate how access to primary neuroimaging data, leading-edge scientific workflow technology, and large-scale computing systems can be performed by any interested researcher to perform a variety of common neuroimaging processing tasks and analyses. We discuss 1) accessing the NDAR data archive, 2) getting data processing workflow software, and 3) how to leverage these to create processing “Pipelines”. Using the

software in question, these workflows can be easily edited, combined, and organized to provide complete end-to-end neuroimaging data processing solutions. Through these basic exemplars, potential users should be able to become familiar with NDAR, workflow design, and their use on remotely accessible high-performance "cloud" computing systems.

Accessing the NDAR Database

In order to access the neuroimaging data stored in NDAR's repository, one must first acquire a username and password by completing the NDAR access application process. This requires completing their user application form, obtaining authorization from an NIH-approved signing official at one's local institution, and submitting the information for formal review by the NIH. The complete details for how to obtain an NDAR account can be found on the NDAR website (http://ndar.nih.gov/ndarpublicweb/access.html).

Upon acquiring an account and logging into the NDAR website, one can begin by using the available "Query" tool in order to filter the collection of data by data types, ages, gender, or number of subjects, etc. Once the results have become sufficiently refined, clicking the "Download Data" button will allow the user to select the types of data they wish to download for their cohort. These data must then be used to create a package, using the "Create Package" button on the right-hand side. Once clicked, it will ask the user to specify name for the package. The NDAR system generates a package ID which is used to identify and refer to that collection of data. The NDAR Download Manager is a program written in Java which will download the collection to one's local hard disk. This method enables users to have data delivered to the computer they are presently using to view the NDAR contents. Rather, users may prefer to reference these NDAR package IDs directly within data processing workflow tools in order to avoid having data directed to remote processing operations located on remote computing systems, clusters, or "in the cloud" using NDAR's "Mini-NDAR" or "miNDAR" capabilities (http://ndar.nih.gov/cloud_get_started.html). Below, we illustrate this process.

Accessing the LONI Pipeline

Scientific workflow methodologies serve to enable the creation of heterogeneous processing chains which then can be run on parallel computing systems. Several such workflow systems exist and have been used successfully in neuroimaging applications (Stef-Praun, Clifford et al. 2007, Gorgolewski, Burns et al. 2011). In particular, the LONI Pipeline workflow environment is a graphical framework for constructing workflows and executing complex high-throughput analysis (Dinov, Van Horn et al. 2009, Dinov, Lozev et al. 2010). This program, now in version 5.x, is freely available (http://pipeline.loni.ucla.edu) and provides processing modules from well-known neuroimaging software programs, as well as complete end-to-end protocols for performing numerous image processing tasks. The use of LONI Pipeline can be employed to help standardize processing methodologies within a laboratory or between research groups supporting the accurate recording of data processing provenance (MacKenzie-Graham, Payan et al. 2008, Mackenzie-Graham, Van Horn et al. 2008), a feature that has been notoriously lacking in the neuroimaging community (Kennedy 2012). LONI Pipeline is also available through the Neuroimaging Tools and Resource

Clearinghouse (NITRC; http://www.nitrc.org) (Luo, Kennedy et al. 2009) – a service enabling researchers to locate, install, and compare resources for functional and structural neuroimaging analyses, as well as collects and points to standardized information about tools for performing such analyses. Users interact with the LONI Pipeline client on PC, Mac, or Linux to design and execute workflows which are actually run through connection to a Pipeline-enabled computer cluster. LONI, itself, maintains a large-scale cluster having 3,500 compute nodes dedicated to supporting thousands of simultaneous Pipeline workflow submissions. Pipeline is also available as an Amazon-EC2 service (http://pipeline.loni.usc.edu/products-services/pipeline-server-on-ec2/) and is also available for use via the NITRC Compute Environment (NITRC-CE) (http://www.nitrc.org/projects/nitrc_es).

Access to these LONI Pipeline–NDAR Example Workflows

With access to LONI Pipeline and a connection to a Pipeline-enabled server, users can easily locate the complete set of example NDAR data processing workflows described here contained within the LONI Pipeline's built-in server library. The workflows are available as supplemental information from this article and are also available for download to a user's local machine via the LONI Pipeline's entry on the NITRC website (http://www.nitrc.org/docman/view.php/32/1294/NDAR%20Workflows%2014Oct2013.zip). We intend that, very soon, users will be able to access these LONI Pipeline workflows directly from within NDAR itself.

NDAR Package ID Access from within LONI Pipeline

Within the Pipeline program there exists the means for a direct login to the NDAR cloud storage and one's previously defined data package IDs (see "Accessing to the NDAR Database," above). Once an NDAR data package is specified in the NDAR login (accessible through the database and cloud storage login interface), a three-component set of modules is automatically generated as a new LONI Pipeline workflow. These modules serve to download the compressed data package, unzip its contents, and convert them to any one of three commonly utilized neuroimaging file formats: Analyze, NIfTI, or MINC. Depending on the user's needs, any of these file types may be selected. With these modules in place, new workflows can then be constructed by connecting them to other processing modules drawn from Pipeline's library of available processing modules and workflows. Processing operations are available from a large selection of commonly used neuroimaging data processing packages including FSL, FreeSurfer, BrainSuite, and Diffusion Toolkit. Additionally, modules defined for user-built processing executables can also be inserted to generate unique workflows via custom data processing toolkits.

Example Pipeline Workflows for NDAR

Here, we seek to illustrate several LONI Pipeline workflows which focus on NDAR data. Our intent here is to provide the context for how the interested user can use these workflows to begin a process of using NDAR data, performing basic processing operations on them, and to set the stage for more all-encompassing analyses using greater numbers of subjects, variables of interest, etc. All workflows have been developed to be straightforward; they use the commonly available and widely used tools mentioned above, depend on only a few

instances where custom tools have been included, and are color-coded in such a way that a user can understand the purpose of the workflow at a glance and easily alter the input data to their own NDAR data package, augment the available processing parameters, and otherwise have a description of what each workflow is designed to do.

Several of the workflows utilize the automatically generated NDAR cloud database retrieval modules discussed above. Once logged-in, users can specify the ID of the NDAR package they have created online, follow annotated instructions on the Pipeline workflows for adjusting a few basic settings, and then click on the LONI Pipeline's green "Start" button to launch the processing job. However, depending on their intentions, the user may find it desirable to download the package of data directly via the NDAR Download Manager, for data inspection and subsequent processing of cases individually, in batches, or based on data type. Note, however, that due to constraints from the NDAR download server, one limitation is that NDAR package numbers cannot be accessed by multiple active workflows simultaneously without using the miNDAR option. Regardless of the means for NDAR data download, upon completion of the full workflow, the resulting files from processing or statistical analysis will be written to the user's specified output directory.

Basic NDAR File Download using Pipeline

The first workflow in the NDAR workflow library is the basic NDAR file acquisition protocol. It features a data source (gray disk), a data sink (inverted triangle), and two modules in between (the purple circles). A complete description of LONI Pipeline iconography, its intentions and functions can be found in the online LONI Pipeline User Handbook (http://pipeline.loni.usc.edu/learn/handbook/). The data source is where the user specifies the ID of a package they have created. The first module, "NDARGet", simply connects to the NDAR webserver to access the package and downloads its files as a .zip file archive. Subsequently, the following module converts the data to the desired user-specified format. The default file type is the NIfTI (.nii) file format (http://nifti.nimh.nih.gov/), though other formats can be selected by the user. Finally, the user needs to specify a location for the data to be deposited in the data sink module. This may be in a local directory on the user's personal computer or on the remote server to which they are connected and where they may access it at a later time.

Brain Extraction

Building on the above workflow, the second workflow performs a very typical data processing "first step" – that of removing the brain from the skull and surrounding non-brain tissues. Following download and .nii conversion, the workflow utilizes FSL's Brain Extraction Toolkit (BET) workflow module which performs the familiar skullstripping of the image volumes. By right-clicking on the BET module, the user can also set the module parameters to estimate the external skull surface, use alternative image thresholding, and adjust the aspect ratio of the volume bounding box. These images are then written to the output directory specified in the data sink, as previously described. Any number of subjects may be entered as inputs to be skull-stripped simultaneously.

FreeSurfer Surface Generator

Researchers frequently wish to create cortical surface models of the subjects in their samples to use for further analysis of brain surface morphometry or for use in the display of functional results. Cortical surface analyses have been of particular interest to ASD researchers (Shokouhi, Williams et al. 2012, Doyle-Thomas, Kushki et al. 2013, Wallace, Robustelli et al. 2013). This workflow allows the user to utilize a LONI Pipeline-enabled FreeSurfer "ReconAll" module to generate two surface object files (.obj) delineating the left and right hemispheres for a single subject or multiple subjects. It utilizes the basic NDAR File Acquisition procedure, passing .nii converted files to the FreeSurfer module, before performing octahedral resampling, and writing separate files for each hemisphere. The user needs to specify an output directory for Pipeline to create prior to running the workflow into which Pipeline will write its results files. This can be a directory on one's local machine or a directory on the filesystem of the Pipeline server being utilized. Modifications to the FreeSurfer module will allow for the creation of a greater number of surface models, as well as detailed labelmaps.

Basic DTI Preprocessing

Recent work with diffusion imaging has indicated that white matter abnormalities may play a role in differentiating ASD phenotypes (Billeci, Calderoni et al. 2012, Delmonte, Gallagher et al. 2013). One particularly useful DTI imaging analysis package is the Diffusion Toolkit (DTK) (http://trackvis.org/dtk/), which provides software designed to reconstruct white matter fiber pathways from DTI image gradient images. The workflow begins, once again, with the basic NDAR File Acquisition procedure for one or more subjects, directing the converted .nii files through DTK reconstruction to output many computational water-molecule diffusion variables such as average diffusion coefficient (ADC), fractional anisotropy (FA), mean diffusivity (MD), and a b0 image which can be used as a structural reference volume during tractography analysis. Next, it runs the reconstructed fibers through the DTK tracking algorithm. DTI gradient information is provided as an external ASCII text file, which provides users a degree of flexibility for easy modification or alteration, if they see fit (e.g. to compare "raw" gradient information against motion corrected versions). The fiber tracking algorithm employs the default FACT algorithm but may be easily modified by the user by choosing an alternative method delineated by the module. Lastly, the fibers are smoothed by the DTK spline filter module before being saved as .trk files to the desired destination. The resulting .trk files can then be viewed and analyzed exclusively in the freely available TrackVis program (http://trackvis.org).

Basic Task-Based fMRI Preprocessing

One key point of interest for understanding cognitive processing in patients diagnosed with an ASD has been the application of task-based functional imaging (Cody, Pelphrey et al. 2002, Pelphrey, Morris et al. 2005, Pelphrey, Morris et al. 2007). In the analysis of fMRI data, the basic preprocessing workflow for a first-level analysis of activation data is typically performed one subject at a time. We demonstrate this process here using an example subject from NDAR. Through Pipeline, the user must first download the desired

data from NDAR using either the Basic NDAR File Acquisition workflow or through the .jar file obtained using the NDAR Download Manager. To illustrate that a range of file formats can be accommodated, downloaded Analyze (e.g. *.img and accompanying *.hdr file) data can then be referenced in the input module of this fMRI preprocessing workflow. A separate file must be listed for each time point the user wishes to analyze for the subject. In addition, an FSL ".fsf" file must first be created through the FSL FEAT interface. An example template .fsf file is also included in the workflow which can be edited in lieu of creating a new .fsf file. A skullstripped structural T1-weighted image is required as well, which, in this example was pre-computed and the resulting stripped image provided using a Pipeline input module. This workflow takes these inputs, merges the many time points, registers them to the MNI 2mm standard atlas space, and updates the .fsf file. Next, it passes this .fsf file to an FSL FEAT module for analysis, outputting the first level statistical analysis of the data in the familiar HTML format into the specified output directory.

FSL Melodic fMRI Workflow

Alterations of resting-state connectivity may underlie aspects of the social interaction deficits observed in certain forms of ASD (Assaf, Jagannathan et al. 2010, von dem Hagen, Stoyanova et al. 2013). As such, independent component analysis (ICA) is commonly used to identify patterns of correlated voxels which reflect "default mode" and other functional networks that may be affected in ASD. Probabilistic independent components analysis (PICA) (Beckmann, DeLuca et al. 2005) can be performed on fMRI data with the FSL Melodic fMRI Workflow. The entire list of time point images in the resting state scan is provided by the user, along with the subject's T1 anatomical image volume, and a reference spatial brain atlas (the default is the MNI Atlas). The workflow performs the steps of a linear registration of the data, calculates an average resting state image, and runs the FSL Melodic analysis.

ROI Extraction Workflow

As an alternative to the use of FreeSurfer, regions-of-interest (ROIs) can also be easily created for anatomical volumes using an example of the ROI Extraction Workflow presented here. This workflow uses the Brain Parser algorithm (Tu and Toga 2007, Tu, Zheng et al. 2007) which applies a Boost learning algorithm to fit predetermined atlas-based regional parcellations to an individual's T1-weighted anatomical volume. Images are first reoriented, bias field corrected, and skullstripped in accordance with the ICBM Atlas space conventions (Mazziotta, Toga et al. 2001, Shattuck, Mirza et al. 2007) using a series of AIR modules (Woods, Grafton et al. 1998, Woods, Grafton et al. 1998). FSL FLIRT is employed for linear registration in advance of performing 3-D spline filtering. Prepared data are then processed using the Brain Parser module. Regional metrics are them computed by leveraging FreeSurfer which are then compiled into an ASCII text file table for later export and offline analyses.

BrainSuite and FreeSurfer Comparison

It is often of interest to compare processing methodologies against one another to examine algorithmic efficiency, accuracy, and other performance factors as well as conduct direct comparisons of processing output. This workflow presents just such a scenario.

As in the other examples discussed here, this workflow begins with the Basic NDAR File Acquisition procedure for one or more subjects, which, in this instance, obtains the files in Analyze format. The converted Analyze images downloaded from NDAR are distributed across three separate workflow modules. The first of these modules, the BrainSuite Cortical Surface Extraction grouped module (Shattuck and Leahy 2002), bias corrects, skull strips, and then performs all of the extraction steps involved in Brain Suite processing, using an ICBM Atlas and BrainSuite labeling conventions.

The second module which employs the converted files from the NDAR package is the Brain Parser (described above) “grouped” module which reorients, bias corrects, skull strips, creates parcellations, runs FSL FLIRT, and applies an affine transform using an ICBM atlas and the Brain Parser directory of pre-determined models.

The hemispheric surface outputs of these two module groups are then automatically translated to the file formats read by the ShapeToolsIO library (http://www.nitrc.org/projects/shapetools/). The module used for this translation is essentially a wrapper for the various ShapePreparer classes provided by the ShapeToolsIO library file format writers. The volume voxel values of these images are then mapped onto the vertices of a surface mesh so that 3-D models of the left and right hemispheres are created. These can be viewed and analyzed in the BrainSuite program (http://www.nitrc.org/projects/brainsuite).

The third and final module that employs the converted files from the NDAR package is the FreeSurfer reconstruction module, just as was performed using the FreeSurfer Surface Generator workflow, above. This utilizes the FreeSurfer ReconAll reconstruction algorithm to generate two pial surfaces for octahedral resampling. These outputs are then subjected to the same translation and surface-mapping modules as the BrainSuite outputs. The resulting images and mesh surfaces from each sub-workflow can be examined, compared, and contrasted to explore overall program performance, or serve as a prelude to the combination of results via a surface modeling meta-workflows (Rex 2004, Leung, Parker et al. 2008).

In general, each of the workflows included within LONI Pipeline in the NDAR library has a specific application each of which forms a typical neuroimaging processing operation that a neuroimaging researcher might be interested in applying. These individual workflows could be combined in a number of different ways to perform end-to-end processing of NDAR data or used to compare the speed, reliability, or accuracy of competing processing steps – for instance, to examine the bias-variance trade-offs associated with different image processing philosophies (Strother, Anderson et al. 2002, LaConte, Anderson et al. 2003).

Discussion

Despite discussions over the years of its many advantages (Van Horn and Gazzaniga 2002), neuroimaging data sharing remains relatively uncommon (Kennedy 2012), the shared data is often only subject to the analysis of its collectors (Van Horn and Gazzaniga 2012). Many researchers feel that the effort to prepare data for sharing conflicts with the time they have to actually perform research, or fear that others may achieve more recognition for results derived from data over which they feel ownership. Others feel that the specific protocols

they use are superior to the formalized standards data sharing requires (Van Horn and Ball 2008). Some worry that the intricacies of their data will not be understood by those who may wish to use it (Teeters, Harris et al. 2008). Since the collection of neuroimaging data is not always orderly, retrospectively preparing data to be archived or shared can compound errors due to the passage of time (Kennedy 2012).

Yet, through data repository models, such as NDAR, OpenfMRI, the LONI IDA, and other approaches, a number of these worries are being addressed. Sharing data that has been carefully collected with distribution in mind should take next to no time if it is uploaded as the data is collected, and if time-consuming processes such as quality checking and anonymizing are performed by a data coordinating center. Under NDAR, for instance, raw data is initially quarantined, and only made available to the community after the collection site has had an opportunity to publish their results. Post processed data are released at publication. Preprocessed data – which includes structural images – is not quarantined although raw/preprocessed data is expected ongoing and analyzed data at time of publication. Thus, fears that contributing authors will be "scooped" by others who download and use their data are, therefore, greatly reduced. Existing and new publications can then be linked to the data online, so that future researchers can understand the full extent of the data collection protocol. Only raw data is expected to be initially shared, so no intellectual property is being given up in this process. Furthermore, NDAR and similar archives are working to establish protocols for citing the use of its data so that researchers get credit for their efforts even when another researcher builds off of their work (http://ndar.nih.gov/data_from_papers.html). Effective use of data sharing has, in fact, increased the visibility of research. For example, contributors to the fMRI Data Center (fMRIDC), saw their work cited widely which garnered new opportunities for collaboration on the basis of their shared data sets (Van Horn and Ishai 2007). The shared NDAR data repository and the LONI Pipeline interactivity can also foster such cooperation in analyzing and drawing conclusions from large amounts of data. Other projects such as the 1000 Functional Connectomes (Milham 2012) and the OpenfMRI Project (Poldrack, Barch et al. 2013) seek to also support and encourage a greater appreciation for what can be scientifically accomplished through the greater availability of shared neuroimaging data. Any data obtained through these resources would be amenable to processing and analysis using tools such as the LONI Pipeline to create useful end-to-end workflows. Thus, sharing data and having it used by others can be advantageous since recent work suggests that data re-use helps to improves researcher citation rates (Piwowar, Day et al. 2007).

Analyzing data collected by others can traditionally be difficult for a number of reasons. Beyond the differences in imaging conventions that vary according to the commercial vendor of the equipment, there is little agreement within the neuroimaging research community on how to collect, organize, and analyze such diverse data (Poline, Breeze et al. 2012). If researchers hope to compare data from different collection sites, every aspect of the data collection protocol must be standardized. This standardization leads to much finer detail in the protocol documentation, which means there is much less difficulty for end users who wish to carefully select comparable data. The uniformity of LONI Pipeline processing is a first step toward standardized data analysis (Mackenzie-Graham, Van Horn et al. 2008).

Allowing researchers to process the same data in the same manner permits more accurate characterization of the relevance of a publication to the current literature and eliminates many common confounding variables. It also makes replicating the results from another site simple and financially feasible for perhaps the first time. Furthermore, it allows a wider variety of researchers to analyze multimodal data; for example, if an investigator who is relatively unfamiliar with DTI, but frequently works with fMRI, he or she may decide that they would rather spend their time analyzing more fMRI data as opposed to learning how to reconstruct and analyze DTI data. With Pipeline, the researcher would be able to string together software modules, without having to learn to run each of these programs from a command line. This, in turn, makes it possible for specialists to include other imaging modalities in their research. Making use of multi-scaled analyses may prove vital in solving complex questions, such as whether structural correlates exist for functional relationships.

LONI Pipeline's interactivity with the NDAR database and availability via NITRC provides an opportunity to many researchers who would otherwise be unable to conduct ASD research on this scale by allowing large amounts of expensive data or longitudinal data to be freely accessed and processed in a timely fashion. What is more, NITRC has recently begun to provide enhanced services such as virtual computing and data storage (http://tinyurl.com/q2nv3js). It is also in the process of broadening its data domains to include MEG/EEG, optical imaging, digital atlasing, genetic imaging, clinical neuroinformatics, computational neuroscience, electrophysiology, computational neuroscience, and neuroimaging genomics and genetics. Examining disorders such as ASDs across the numerous modalities of data offered by NDAR using Pipeline workflows and NITRC compute resources would enable a broader community of researchers the ability to obtain a more complete picture of the factors at play in the complicated neurological questions.

NDAR also utilizes link-out capability within PubMed and is now issuing Digital Object Identifiers for shared studies, so that a simple search can lead researchers to the actual data used in their references. This has vast potential for comparative research, or for following up on the future areas of research that an author suggests in their discussion. In this way, NDAR interactivity with LONI Pipeline can help serve as a hypothesis generator and fill knowledge gaps in our understanding of ASD. Conversely, within the NDAR website, one can search for data by the lab that the study is associated with and be linked directly to all of the publications associated with a specific cohort. In this way, inquisitive researchers with an interest in autism can see how the data has been used in the past, which may help them shape hypotheses for future study and analysis using Pipeline.

Indeed, large-scale analysis performed using and contrasting true populations of subjects, from multiple sites, and regions of the ASD spectrum are within reach. The example workflows featured here have been constructed to illustrate the processing of data from single individuals. This was done to ensure that they can be run and finish in a reasonable amount of time for the purposes of demonstration when run by interested readers. However, where possible, most can be easily extended to accommodate neuroimaging data from multiple, in some instances hundreds of, cases. Indeed, the total number of subjects which can be processed is limited only by the available data, available computing capacity, and the imagination of the researcher.

Collaborators who have written their own programs, particularly programs created for command line execution, can easily create LONI Pipeline modules to call these programs and thereby can process large swaths of subjects in the same amount of time that it previously took them to process a single subject. Additionally, other researchers may be able to use these new modules to process data available through NDAR to get a better understanding of aspects of autism for which they may not have collected data, but which coincides with their current research. Such workflows can easily be included as supplemental information in publications, or e-mailed to collaborators to standardize procedures of analysis.

Conclusion

Overall, newly developed interactivity between the NDAR database and LONI Pipeline combine the well-documented benefits of data sharing with efficient and streamlined image processing workflows. This opens up data and processing methods from the field of ASD research to many other research centers which would previously have been unable to participate in the community, due to budgetary restrictions, lack of experience, lack of access to large cohorts, or computational resources. Comparative and longitudinal studies will be able to be performed instantly using accumulated data collected, leading to the generation of new, testable hypotheses, and, therefore, new data collection and research innovations. Furthermore, the secondary validation of data, and ease of replication will increase the reliability of the literature base as a whole. Through this linkage of ASD neuroimaging data and leading-edge workflow technologies, we anticipate enriching the study of brain biomarkers by the broadest possible set of interested researchers.

Acknowledgments

This work was supported by the National Institutes of Health grant “Multimodal Developmental Neurogenetics of Females with ASD” (5R01MH100028-03) to K.P. and its sub-award to J. D. V. H. In addition, NIH grants MH083320 and HD004147 support D.N.K. and C.H. We wish to thank the NDAR staff assisted in providing information, data and techniques that contributed to this publication. Finally, we acknowledge the dedicated staff of the Institute for Neuroimaging and Informatics at the University of Southern California.

References

Assaf M, Jagannathan K, Calhoun VD, et al. Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. Neuroimage. 2010; 53(1):247–256. [PubMed: 20621638]

Atkinson J, Braddick O. From genes to brain development to phenotypic behavior: “dorsal-stream vulnerability” in relation to spatial cognition, attention, and planning of actions in Williams syndrome (WS) and other developmental disorders. Prog Brain Res. 2011; 189:261–283. [PubMed: 21489394]

Beckmann CF, DeLuca M, Devlin JT, et al. Investigations into resting-state connectivity using independent component analysis. Philos Trans R Soc Lond B Biol Sci. 2005; 360(1457):1001–1013. [PubMed: 16087444]

Billeci L, Calderoni S, Tosetti M, et al. White matter connectivity in children with autism spectrum disorders: a tract-based spatial statistics study. BMC Neurol. 2012; 12:148. [PubMed: 23194030]

Breeze JL, Poline JB, Kennedy DN. Data sharing and publishing in the field of neuroimaging. Gigascience. 2012; 1(1):9. [PubMed: 23587272]

Cody H, Pelphrey K, Piven J. Structural and functional magnetic resonance imaging of autism. Int J Dev Neurosci. 2002; 20(3–5):421–438. [PubMed: 12175882]

Delmonte S, Gallagher L, O'Hanlon E, et al. Functional and structural connectivity of frontostriatal circuitry in Autism Spectrum Disorder. Front Hum Neurosci. 2013; 7:430. [PubMed: 23964221]

Dinov I, Lozev K, Petrosyan P, et al. Neuroimaging study designs, computational analyses and data provenance using the LONI pipeline. PLoS One. 2010; 5(9)

Dinov ID, Van Horn JD, Lozev KM, et al. Efficient, distributed and interactive neuroimaging data analysis using the LONI pipeline. Frontiers in Neuroinformatics. 2009; 3

Doyle-Thomas KA, Kushki A, Duerden EG, et al. The effect of diagnosis, age, and symptom severity on cortical surface area in the cingulate cortex and insula in autism spectrum disorders. J Child Neurol. 2013; 28(6):732–739. [PubMed: 22832774]

Gorgolewski K, Burns CD, Madison C, et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front Neuroinform. 2011; 5:13. [PubMed: 21897815]

Hall D, Huerta MF, McAuliffe MJ, et al. Sharing heterogeneous data: the national database for autism research. Neuroinformatics. 2012; 10(4):331–339. [PubMed: 22622767]

Insel TR, Volkow ND, Landis SC, et al. Limits to growth: why neuroscience needs large-scale science. Nat Neurosci. 2004; 7(5):426–427. [PubMed: 15114352]

Kennedy DN. The benefits of preparing data for sharing even when you don't. Neuroinformatics. 2012; 10(3):223–224. [PubMed: 22661300]

LaConte S, Anderson J, Muley S, et al. The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics. Neuroimage. 2003; 18(1):10–27. [PubMed: 12507440]

Leung K, Parker DS, Cunha A, et al. IRMA: an Image Registration Meta-Algorithm - evaluating Alternative Algorithms with Multiple Metrics. SSDBM. 2008

Luo, X-zJ; Kennedy, DN.; Cohen, Z. Neuroimaging informatics tools and resources clearinghouse (NITRC) resource announcement. Neuroinformatics. 2009; 7(1):55–56. [PubMed: 19184562]

MacKenzie-Graham, A.; Payan, A.; Dinov, I., et al. Provenance and Annotation of Data International Provenance and Annotation Workshop, IPAW 2008. Salt Lake City, UT: University of Utah; 2008. Neuroimaging Data Provenance Using the LONI Pipeline Workflow Environment.

Mackenzie-Graham AJ, Van Horn JD, Woods RP, et al. Provenance in neuroimaging. Neuroimage. 2008; 42(1):178–195. [PubMed: 18519166]

Mazziotta J, Toga A, Evans A, et al. A four-dimensional probabilistic atlas of the human brain. J Am Med Inform Assoc. 2001; 8(5):401–430. [PubMed: 11522763]

McPartland JC, Coffman M, Pelphrey KA. Recent advances in understanding the neural bases of autism spectrum disorder. Curr Opin Pediatr. 2011; 23(6):628–632. [PubMed: 21970830]

Milham MP. Open neuroscience solutions for the connectome-wide association era. Neuron. 2012; 73(2):214–218. [PubMed: 22284177]

National Autism Center. Evidence-based Practice and Autism in the Schools. Randolph; Massachusetts: 2011.

Pelphrey KA, Morris JP, McCarthy G. Neural basis of eye gaze processing deficits in autism. Brain. 2005; 128(Pt 5):1038–1048. [PubMed: 15758039]

Pelphrey KA, Morris JP, McCarthy G, et al. Perception of dynamic changes in facial affect and identity in autism. Soc Cogn Affect Neurosci. 2007; 2(2):140–149. [PubMed: 18174910]

Piwowar HA, Day RS, Fridsma DB. Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE. 2007; 2(3):e308. [PubMed: 17375194]

Poldrack RA, Barch DM, Mitchell JP, et al. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform. 2013; 7:12. [PubMed: 23847528]

Poline JB, Breeze JL, Ghosh S, et al. Data sharing in neuroimaging research. Front Neuroinform. 2012; 6:9. [PubMed: 22493576]

Rex DE, Shattuck DW, Woods RP, Narr KL, Luders E, Rehm K, Stolzner SE, Rottenberg DE, Toga AW. A meta-algorithm for brain extraction in MRI. Neuroimage. 2004; 23(2):625–637. [PubMed: 15488412]

Shattuck DW, Leahy RM. BrainSuite: an automated cortical surface identification tool. Med Image Anal. 2002; 6(2):129–142. [PubMed: 12045000]

Shattuck DW, Mirza M, Adisetiyo V, et al. Construction of a 3D probabilistic atlas of human cortical structures. Neuroimage. 2007

Shokouhi M, Williams JH, Waiter GD, et al. Changes in the sulcal size associated with autism spectrum disorder revealed by sulcal morphometry. Autism Res. 2012; 5(4):245–252. [PubMed: 22674695]

Stef-Praun T, Clifford B, Foster I, et al. Accelerating Medical Research using the Swift Workflow System. Stud Health Technol Inform. 2007; 126:207–216. [PubMed: 17476063]

Strother SC, Anderson J, Hansen LK, et al. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. Neuroimage. 2002; 15(4):747–771. [PubMed: 11906218]

Teeters JL, Harris KD, Millman KJ, et al. Data sharing for computational neuroscience. Neuroinformatics. 2008; 6(1):47–55. [PubMed: 18259695]

Tu Z, Toga AW. Towards whole brain segmentation by a hybrid model. Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv. 2007; 10(Pt 2):169–177.

Tu Z, Zheng S, Yuille AL, et al. Automated extraction of the cortical sulci based on a supervised learning approach. IEEE Trans Med Imaging. 2007; 26(4):541–552. [PubMed: 17427741]

Van Horn JD, Ball CA. Domain-specific data sharing in neuroscience: what do we have to learn from each other? Neuroinformatics. 2008; 6(2):117–121. [PubMed: 18473189]

Van Horn, JD.; Dobson, J.; Woodward, J., et al. Grid-Based Computing and the Future of Neuroscience Computation. In: Senior, C.; Russell, T.; Gazzaniga, MS., editors. Methods in Mind. Cambridge: MIT Press; 2006. p. 141-170.

Van Horn JD, Gazzaniga MS. Databasing fMRI Studies - Toward a ‘Discovery Science’ of Brain Function. Nature Reviews Neuroscience. 2002; 3(4):314–318.

Van Horn JD, Gazzaniga MS. Why share data? Lessons learned from the fMRIDC. Neuroimage. 2012

Van Horn JD, Ishai A. Mapping the human brain: new insights from FMRI data sharing. Neuroinformatics. 2007; 5(3):146–153. [PubMed: 17917125]

von dem Hagen EA, Stoyanova RS, Baron-Cohen S, et al. Reduced functional connectivity within and between ‘social’ resting state networks in autism spectrum conditions. Soc Cogn Affect Neurosci. 2013; 8(6):694–701. [PubMed: 22563003]

Wallace GL, Robustelli B, Dankner N, et al. Increased gyrification, but comparable surface area in adolescents with autism spectrum disorders. Brain. 2013; 136(Pt 6):1956–1967. [PubMed: 23715094]

Woods RP, Grafton ST, Holmes CJ, et al. Automated image registration: I. General methods and intrasubject, intramodality validation. J Comput Assist Tomogr. 1998; 22(1):139–152. [PubMed: 9448779]

Woods RP, Grafton ST, Watson JD, et al. Automated image registration: II. Intersubject validation of linear and nonlinear models. J Comput Assist Tomogr. 1998; 22(1):153–165. [PubMed: 9448780]

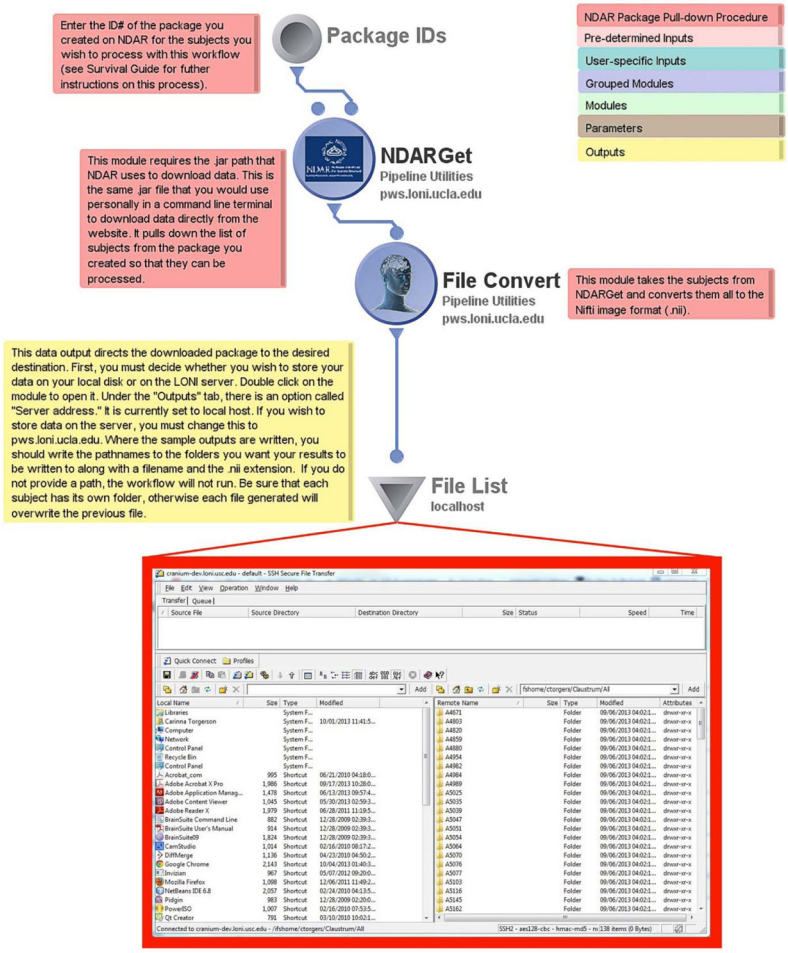

Figure 1.

The Basic NDAR File Acquisition workflow includes annotations that will help the user get acquainted with how they ought to interact with input and output modules and how they can retrieve data from NDAR via the LONI Pipeline environment. The inset picture shows a sample NDAR download file structure for data to be saved to one's local server for further use. Each Pipeline workflow includes the color-coded key, shown at the top right corner of this figure, which allows users to quickly discern which parts of the workflow will require adjustment before processing can begin.

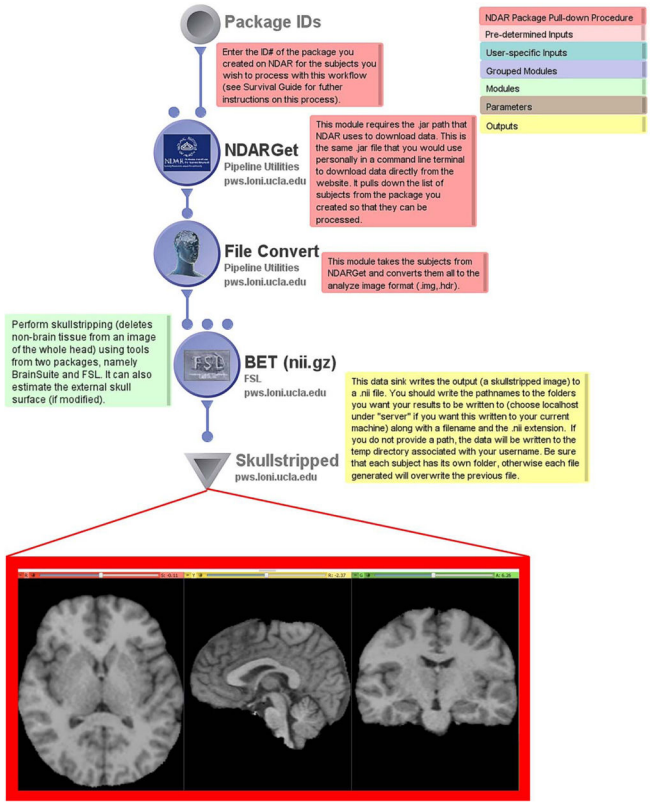

Figure 2.

The Brain Extraction Toolkit workflow outputs a NIfTI volume file, in which non-brain matter has been removed which can then be used in a myriad of other neuroimaging applications or external software packages.

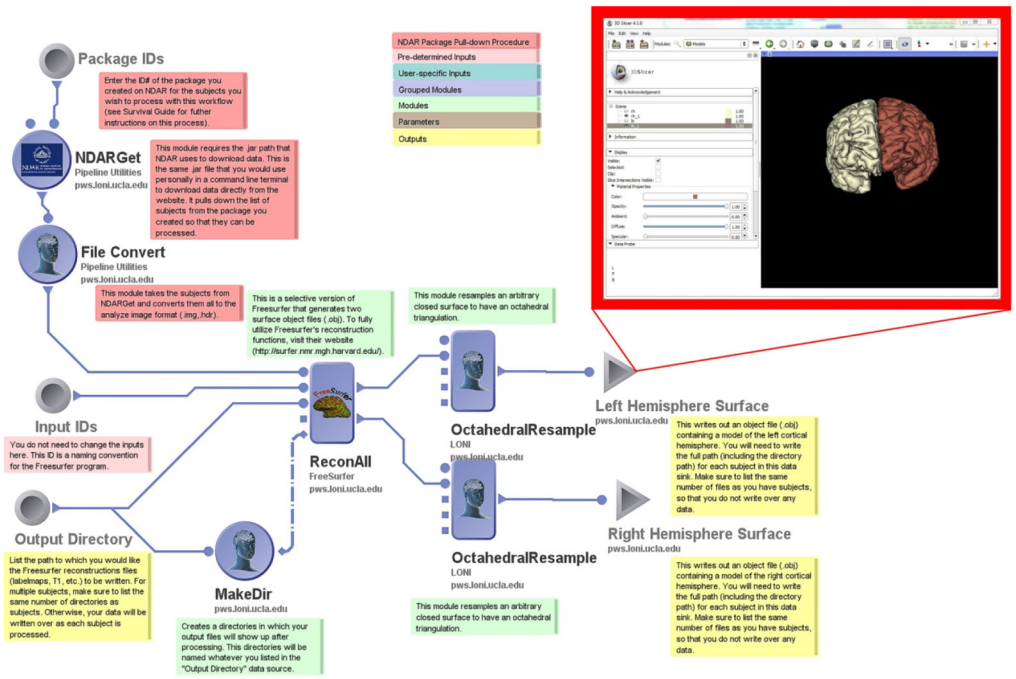

Figure 3.

The FreeSurfer Surface Generator workflow creates object files that contain 3D images of the regions in a FreeSurfer labelmap. In this workflow, only the two grey matter hemispheres are generated, but by editing the ReconAll module in Pipeline, users can generate models of any region present in a labelmap.

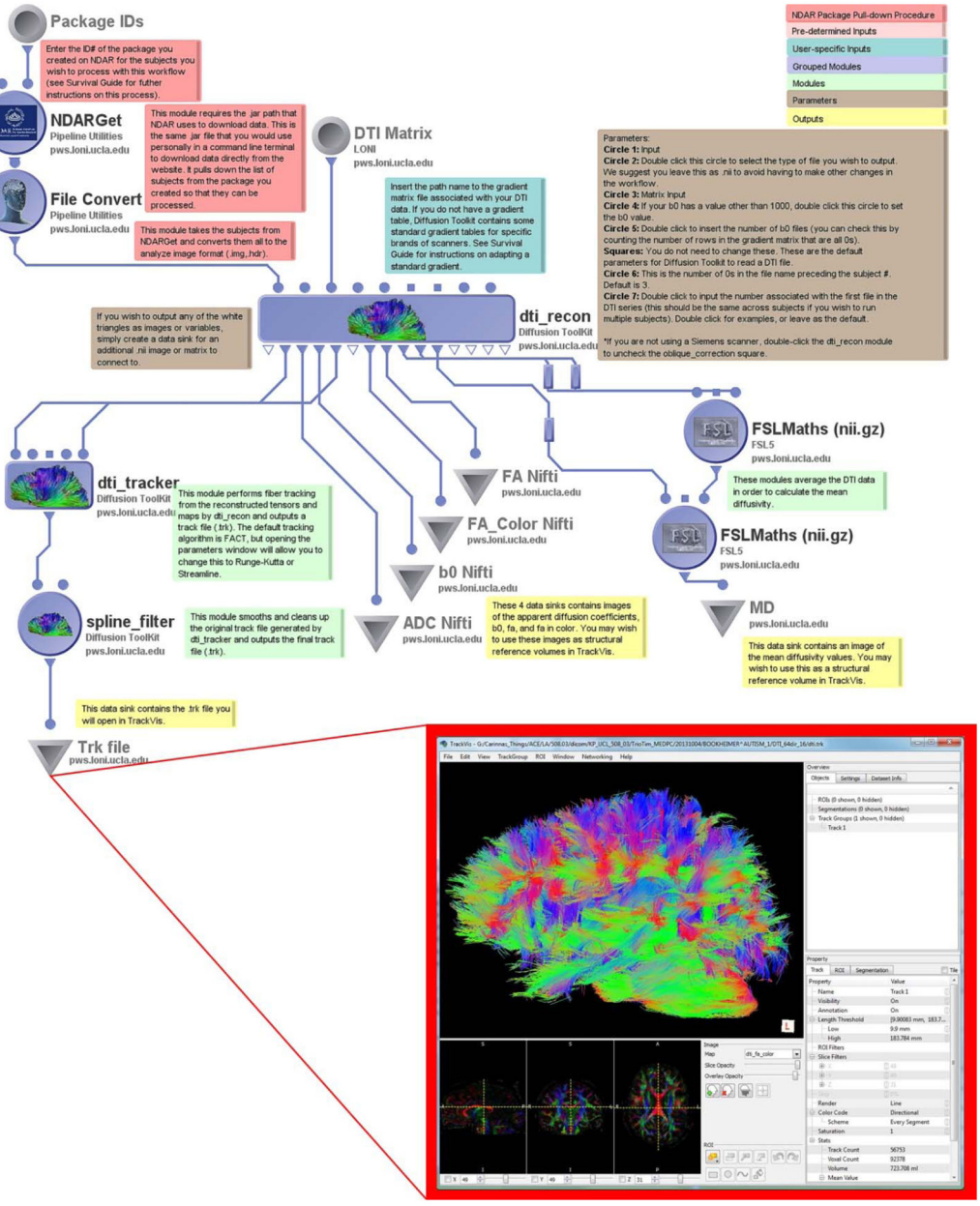

Figure 4.

Tractography files can be quickly generated using the Basic DTI Processing workflow. These .trk files can be viewed, analyzed, and edited in TrackVis, or can be further converted to .vtk files for use in other neuroimaging software such as 3D Slicer (www.slicer.org).

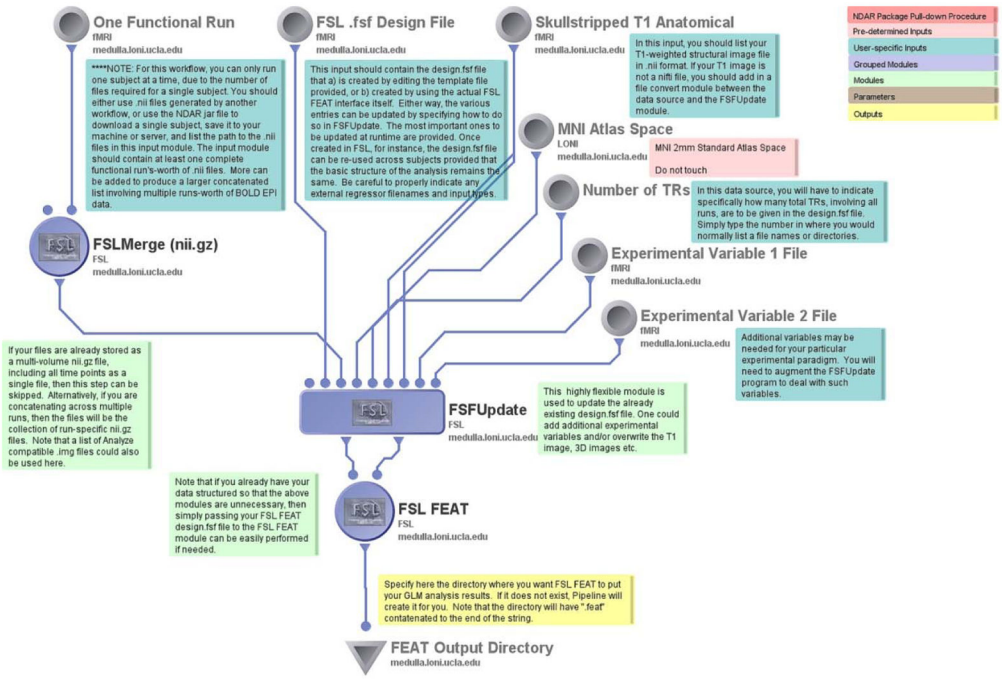

Figure 5.

FSL FEAT, serves as the major processing module in the Basic fMRI Preprocessing workflow, takes as input a pre-defined .fsf file, and then generates a directory containing HTML output, image, and log files of a first level fMRI analysis. The workflow also outputs MATLAB for use in further statistical analysis performed outside of Pipeline if desired.

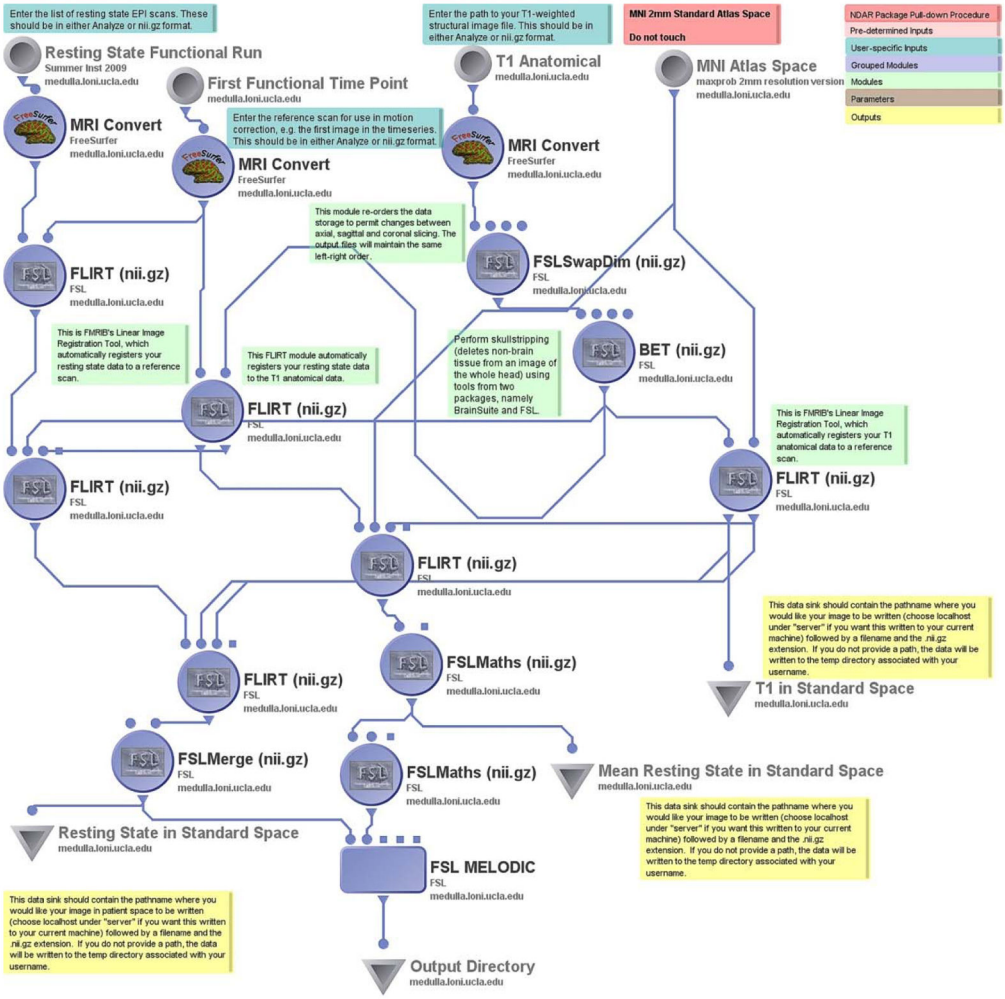

Figure 6.

The FSL Melodic fMRI Workflow generates mean resting state and standardized anatomy image files, as well as a directory with HTML files containing the FSL Melodic report.

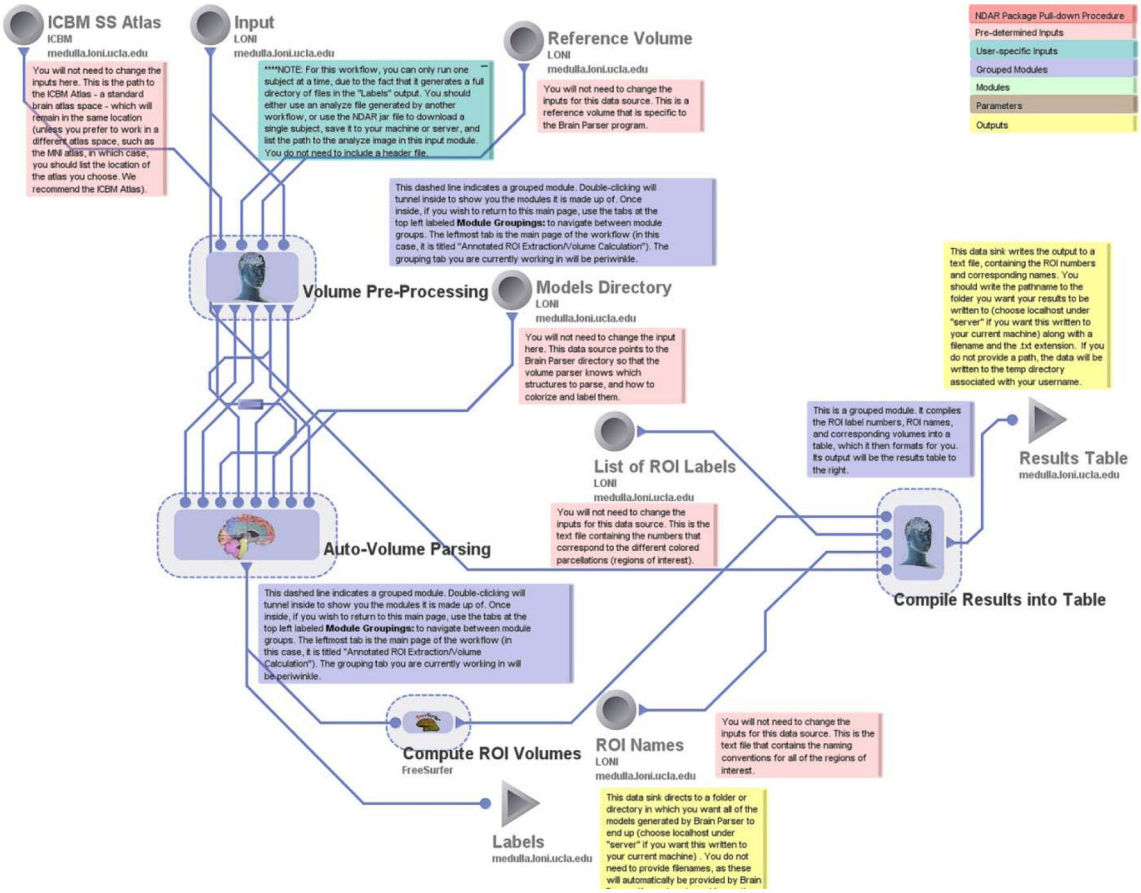

Figure 7.

ROI Extraction parses an anatomical volume into labels and produces a table to help the user identify the names of the regions that the Brain Parser program uses. Models of each label look similar to the inset image in Figure 3.

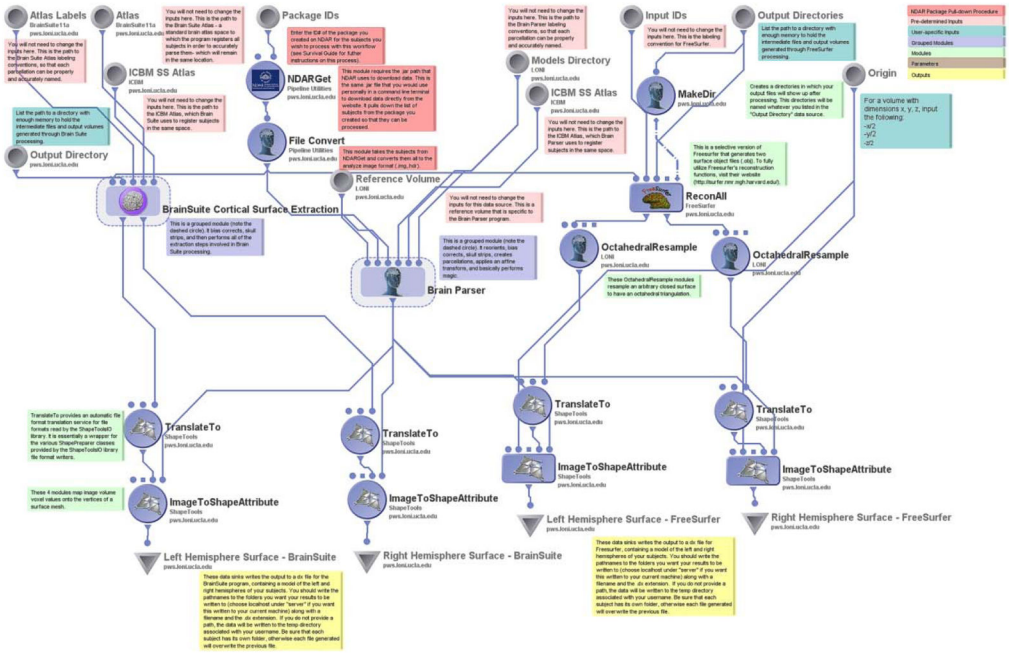

Figure 8.

Brain Suite and FreeSurfer are the most popular neuroimaging software programs for creating 3D models of labelmaps obtained from parsing an anatomical image. The BrainSuite and FreeSurfer Modeling Workflow outputs left and right hemisphere object files in the formats preferred by each of these programs for comparison and examination.