**What is Agua**

Agua is a workflow and genomic visualization platform for running highly scalable high performance computing analyses in the Amazon Elastic Compute Cloud (EC2). Agua features pre-configured analysis pipelines for Next Generation Sequencing (NGS) data and enables users to easily and rapidly create custom analyses. For an overview of Agua, please see the [**Introduction**](http://www.aguadev.org/confluence/display/home/Introduction).



### **Who is it for?**

Agua is about empowering two types of people: **Non-technical Users**who want to run complex, prepackaged analyses without having to set up a bioinformatics/IT department, and **Technical Users**who need to efficiently manage and maintain existing pipelines and save time constructing new pipelines.

**How do it get it?**

Please see the [**Get It**](http://www.aguadev.org/confluence/display/getit/Get+It)section of this website for details on how to start Agua in the EC2 cloud or download and install it on your own host.

**How do I learn more?**

For an overview of Agua, please see the [**Introduction**](http://www.aguadev.org/confluence/display/home/Introduction).

For further information regarding Agua and its functionality, please see [**Frequently Asked Questions**](http://www.aguadev.org/confluence/display/home/FAQ).

And for the latest discussions see the [**Mailing List**](http://www.aguadev.org/confluence/display/home/Mailing+List).

The [**How To**](http://www.aguadev.org/confluence/display/howto/How+To) section describes how to use Agua.

And the [**Developers**](http://www.aguadev.org/confluence/display/dev/Developers) section describes how to modify and contribute to it.

### **What is Agua?**

Agua is a workflow and genomic visualization platform for running highly scalable high performance computing analyses in the Amazon Elastic Compute Cloud (EC2). Agua features pre-configured analysis pipelines for Next Generation Sequencing (NGS) data. Custom analyses can also be easily configured. See the [Introduction](http://www.aguadev.org/confluence/display/agua/Introduction) for more details.

### **Who is it for?**

Agua is about empowering two types of people:

**Non-technical Users** who want to run complex analyses without having to set up a bioinformatics/IT department

**Advanced Users** who need to save time constructing new pipelines and managing/maintaining existing ones

For further information regarding Agua and its functionality, please see [Frequently Asked Questions](http://www.aguadev.org/confluence/display/agua/FAQ). The [How To](http://www.aguadev.org/confluence/display/howto/How+To) section describes how to use Agua and the [Developers](http://www.aguadev.org/confluence/display/dev/Developers)section describes how to modify and contribute to it.

### **How do I get it?**

Please see the [Get It](http://www.aguadev.org/confluence/display/getit/Get+It) section of this website.

## **News Calendar**

[**Wednesday, December 12, 2012**](http://www.aguadev.org/confluence/display/home/2012/12/12)

[**Agua 0.8.0-beta.1 released**](http://www.aguadev.org/confluence/display/home/2012/12/12/Agua+0.8.0-beta.1+released)

Last changed Dec 13, 2012 23:43 by [**Stuart Young**](http://www.aguadev.org/confluence/display/~admin)

# **Agua 0.8.0-beta.1 released**

**Features:**

1. **Drag and drop workflows - use preconfigured workflows or easily create your own custom workflows**
2. **Generate custom on-demand EC2 clusters using**[**StarCluster**](http://star.mit.edu/cluster)**for your specific computing needs**
3. **Real-time visualization of analysis results using**[**JBrowse genomic viewer**](http://www.jbrowse.org/)
4. **Easily view, upload and download your data files using the desktop-style 'Folders' panel**
5. **Multi-user platform allows teams to view, share and copy workflows**
6. **Archive snapshots of your workflows to your**[**GitHub**](http://www.github.com/)**account**
7. **Share/publish a versioned 'biorepository' of your workflows on your**[**GitHub**](http://www.github.com/)**account**

### **Introduction**

The increasingly data-driven nature of the life sciences provides the motivation for harnessing the computational power of cloud computing to tackle important biological questions. Cloud computing allows the user to start up as many computers as needed for an analysis, pay for them by by the hour and then switch them off when the work is done. However, this democratization of computational access is hindered by the technical hurdles of setting up and managing cloud resources. Agua simplifies cloud computing by integrating an intuitive, drag-and-drop web interface with a framework for running workflows on instantly created private clusters within the EC2 cloud.

### **Workflows**

Workflows can be created and run using simple drag and drop controls. Users can create new pipelines using Agua's pre-installed application components and can add their own custom applications to workflows. Custom workflows can be easily created and Agua also enables workflow sharing for improved verification of bioinformatics procedures.

### **Applications**

Users can add new applications easily to the existing suite of applications. These custom applications can be configured to 'recognize' their inputs from the outputs of preceding applications in a workflow. Agua's Input/Output model allows users to fashion coherent and flexible input-output rules for almost any collection of applications.

### **Groups**

Teams of users can share pipelines in the EC2 cloud and manage data access with user-defined groups. Users can flexibly create groups, add members to groups and share data analysis pipelines and results with group members. Multi-level access privileges can be allocated to group members and to users in the general public.

### **Cloud computing**

Agua workflows can be run locally on the head node (where Agua is installed) or on [StarCluster](http://www.stardev.org/) cloud node clusters. Once a cluster workflow is started, the StarCluster begins to auto-scale itself, adding new nodes to the cluster when the number and average duration of jobs is high and removing nodes when demand slacks off. The user can preconfigure the minimum and maximum number of nodes for each StarCluster tailored to each workflow. The user can also change the minimum and maximum number of nodes while the workflow is in progress to best suit their budget and time constraints.

### **Genomic Viewer**

Using Agua's default pipelines, Next Generation Sequencing data can be processed and viewed in the built-in [JBrowse genome viewer](http://www.jbrowse.org/). JBrowse employs the same technology as Google-Maps to allow the user to scan through the genome, zooming in and out in real time.

### **Frequently Asked Questions**

**Can I install a local copy of Agua?**

Agua is designed for use on a computing cluster using a job scheduler: either LSF (Load Sharing Facility), PBS (Portable Batch System) or SGE (Sun Grid Engine), and will also run on a single host. You can [download the source code](http://www.aguadev.org/confluence/pages/createpage.action?spaceKey=agua&title=Download&linkCreation=true&fromPageId=3407884) and run the config.pl file inside the bin/scripts directory to configure the installation step-by-step for single-host or cluster mode.

System requirements are minimal: Apache and MySQL on Linux, approximately 512MB of RAM, 1GB of disk space for Agua and an additional 1. Agua's full functionality is supported on Centos 5.4 but it will also work on other flavors of Linux such as MacOSX and Ubuntu, and will also work on Windows.

**Can I customize the applications available for workflows?**

The administrative user inside the Agua account system can add additional applications that have been made available to the 'aqua' user account on the operating system. During the Agua installation process, you will be prompted for the name of the Administrator account (default: 'admin'). The Administrator can add/remove/edit workflow applications as desired by using the 'Apps' pane of the 'Admin' page.

**What kind of sequences can you use?**   
The most commonly used NGS sequencing platforms are supported: Illumina/Solexa, AB/SOLiD and Roche/454.

**How do I share my data?**  
In the 'Access' tab of the Admin pane, you can configure sharing of your workflows and data with other users.

**What license is Agua released under?**  
All Agua components are open source and free for academic/non-profit organisations.

**What is the maximum amount of data I can have in my Agua account?**

*Your Agua data is stored in the '.Agua' directory inside your CCS home folder. The maximum amount of data you can store in your CCS home folder depends on which group you belong to. Non-MIHG users belong to a CCS group (the Group Name in the*[*New User form*](http://gouda.ccs.miami.edu:8090/display/NextGen/Get+a+CCS+account)*).*

*For non-MIHG users, the standard quota for a group is 1 TB shared among all members of the group, which means that the amount of data you can store depends on how much the other members of your group have in their home folders. Once the total data for a group has exceeded 1 TB (the soft limit), the group has 7 days to reduce the total below 1 TB, after which data will be arbitrarily removed if the total amount exceeds 1 TB. The total data storage for a group cannot exceed 1.1 TB even temporarily (hard limit). You can check your disk usage status by typing on the command line: mmlsquota -g GROUPNAME, where GROUPNAME is the name of your group.*

*All MIHG users are treated as one group and have their home directories in the /mihg/users partition. Storage for the whole group is limited to the available space in that partition.*

**Maximum file size**  
What is the maximum size of data files?

*There is no restriction on the size of data files provided the total data size does not exceed the limits described above.*

**Saved workflows**  
How long will my workflows and data be stored in Agua?

*When you run workflows in Agua, your generated data and the workflow configurations you used are stored to the database. You can recall stored workflows and analysis data and these will be stored in Agua unless you delete them. Your data is stored in your home folder and can only be deleted by you on the command line, or in Agua by yourself or by users who you have allowed 'write' access (on a group-by-group basis) using the 'Access' pane of the 'Admin' page.*

*If you delete a workflow using the 'Delete' menu option in the 'Workflow' page, the workflow and history and all of its associated data will be instantly removed from disk. You then have 30 days within which to retrieve your deleted data from the data backup tape archive, which generates a backup of all home folders every night.*

**Why is it called 'Agua'?**  
It's an acronym: Automated Genomics Utilities Agent

**PURPOSE AND AIMS**

Due to exponentially increasing data volumes from NextGen (next generation) sequencing technologies there is a bioinformatics bottleneck at the analysis stage of biomedical research. There is an urgent need for easy-to-use and widely-accessible tools to accomplish complex bioinformatics tasks that are increasingly composed of multiple interdependent steps. Visualization tools are also needed that are capable of handling the millions of reads produced by NextGen and integrating databases of genomic features.

Figure 1 shows the capabilities and limitations of currently available genome viewers and bioinformatics workflow tools compared to the current and future versions of Agua. Current genome viewers are limited by their technologies. For example, mainstream genome browsers by such as the UCSC Genome Browser, Ensembl Genome Browser and the Generic Genome Browser (GBrowse) employ the traditional thin client-server paradigm with its inherent speed and interactivity constraints. For these browsers, every time the user selects a different view of a genome, the browser must download a new image file from the server so the user experience is highly dependent on Internet connectivity and transfer speeds. Using image files also brings with it implicit constraints on interactivity with the graphical representation and its underlying data set. Alongside these drawbacks, both UCSC and Ensembl provide access to large and diverse sets of genome features which are integrated into the genome view.



**Figure 1. Functionalities of genome browsers and workflow tools**

The Java-based Integrative Genome Viewer (IGV) has slightly improved interactivity but depends on having a Java runtime environment installed on the client. The Java-based genome annotation package, Apollo, also has better interactivity than traditional web page genome browsers but only limited capability for web-based data sharing and access and for integrating external genome feature sets.

AJAX GBrowse breaks with tradition by using asynchronous data transport (where the web page is not reloaded), on-demand loading and nested container lists for quick searching to provide a fast-rendering DHTML display of genomic features. The DHTML paradigm lends itself to much greater interactivity, because it allows access to the DOM (Document Object Model) using Javascript. Furthermore, AJAX GBrowse is based on the open-source and well-supported Dojo Javascript Toolkit which contains many useful tools for rapid web application development.

### **Functional requirements**

The following list of requirements encompasses the functionality envisaged for the workflow and genome view tool, henceforth referred to as Agua (Automated Query and Workflow Agent).  
**Functional Requirements**

1. Low barriers to usage
   1. Web access
   2. Intuitive interface
   3. Plug 'n play applications for rapid deployment
   4. Searchable project annotations
2. Workflows
   1. Predefined workflows (transcriptome, variation annotation, gene networks, file conversion utilities, ID conversion utilities, etc.)
   2. Customizable workflows
   3. Drag 'n drop workflows
   4. Persistent data and workflow configurations
   5. Loops, conditional branching
3. Reports - customizable report extraction from workflow output
4. Views
   1. Customizable genomic views of report data
   2. Interactive display with rich context menu
   3. Integrated, extensible genomic features
   4. Multiple feature views - nucleotide level to aggregate high-level view
   5. Fast view update
   6. User can filter view based on data
   7. Genomic feature-level annotation by user
5. Sharing
   1. User-defined groups with customizable permissions
   2. Workflow, report and view sharing among groups
6. Input/Output and execution
   1. Import external biological data and genomic features into workflow
   2. Integration with external software (e.g., Cytoscape, R, GeneSet Analyzer)
   3. Programmatic remote access (API and Web Service)
   4. Cluster execution
   5. Grid execution
7. Maintain state (action history)
8. Data management - direct user access to input and output files

Information regarding data provenance (Simmhan, 2005) is retained in the system to identify the source of data throughout the workflow such as the owner, author application, creation and modification dates, and content type. In addition, a log is kept of all project changes and updates. The user can also annotate the project at all levels and search these annotations. Agua is similar to the laboratory notebook paradigm employed by the BCJ (Bioinformatics Computational Journal) workflow tools (Feagan, 2007) but has a wider the range of functionality. Agua's functional requirements largely encompass those for a proposed 'genome wiki'9Salzberg, 2007) intended to facilitate cooperative genome annotation by a community of experts, reflecting Agua's utility to the wider genomics community as an accurate, continually updated source of genome annotation. Agua's genome viewer functionality incorporates the AJAX-enabled JBrowse genome viewer ([http://jbrowse.org](http://jbrowse.org/)) to provide a fast, fluid and responsive genome browser interface. 



(Figure 2).

**References**

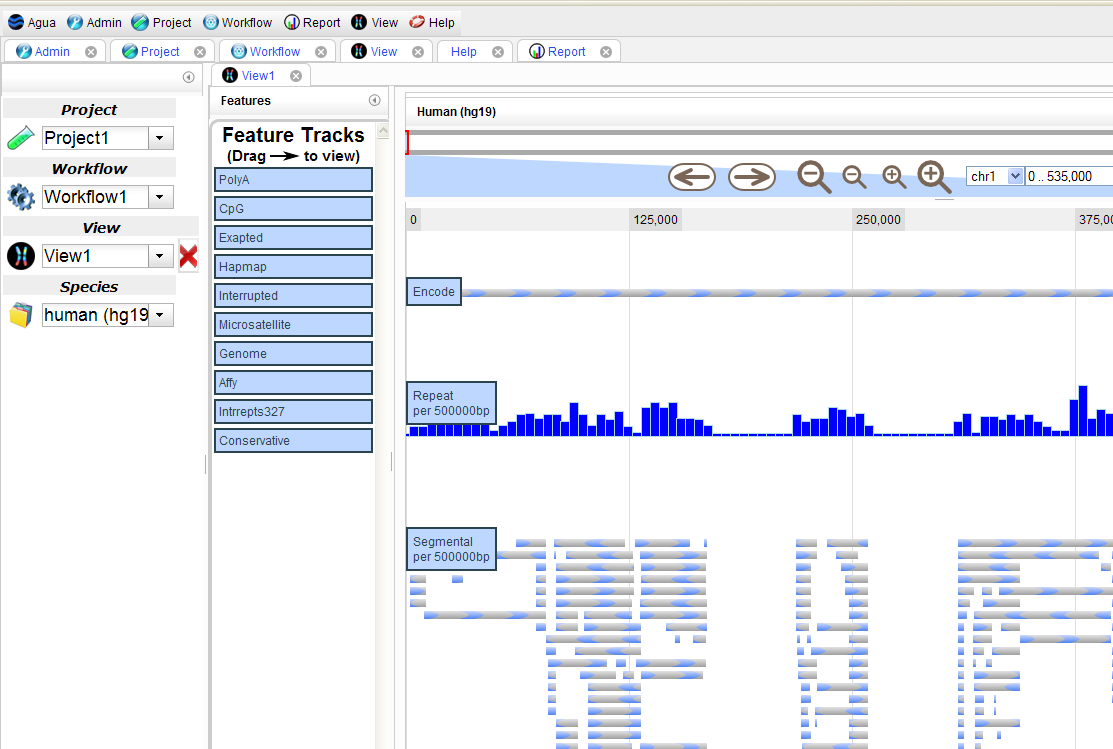
Simmhan Y, Plale B, Gannon D. A survey of data provenance in e-science. SIGMOD Record. 2005;34(3):31 - 6.

Feagan L, Rohrer J, Garrett A, Amthauer H, Komp E, Johnson D, et al. Bioinformatics process management: information flow via a computational journal. Source Code for Biology and Medicine. 2007;2(1):9.

Salzberg S. Genome re-annotation: a wiki solution? Genome Biology. 2007;8(1):102.

**Genomic Viewer**

Agua contains an integrated [JBrowse](http://jbrowse.org/) Javascript-based genomic viewer for visualizing large datasets rapidly and interactively.



# **Reference sequences**

The 'Sources' folders in Agua are located in /nethome/bioinfo/data/sequences so that they are accessible to all users:

/nethome/bioinfo/data/sequence/

|

|

cdna----------+

| human-bfa---+

| | embl

| | ncbi

| human-fa----+

| | embl

| | ncbi

| |

| human-sq----+

| ...

|

chromosomes--+

| human-bfa---+

| | chr1.bfa

| | chr2.bfa

| | ...

| |

| human-fa

| human-sq

|

|

genomes

| human-bfa---+

| | chr1.bfa

| | chr2.bfa

| | ...

| |

| human-fa

| human-sq

|

mtdna---------+

...

# **Get It**

There are two ways to obtain your own personal Agua installation:

## [**1. Launch the Agua AMI in the EC2 cloud (nothing to install)**](http://www.aguadev.org/confluence/display/getit/1.+AMI+Launch+Steps)

This is the simplest way to get up and running with Agua. All you need is a credit card and an internet connection.

Follow [the AMI launch steps](http://www.aguadev.org/confluence/display/getit/1.+AMI+Launch+Steps) to create an EC2 account and launch your own Agua instance.

## [**2. Install Agua on your own server (download and install)**](http://www.aguadev.org/confluence/display/getit/2.+Installation+Steps)

If you want to use Agua on your own custom AMI or want to use it locally to run applications in the cloud (make sure you have enough bandwidth for file transfers), please browse the [Installing Agua](http://www.aguadev.org/confluence/display/getit/2.+Installation+Steps) guidelines. You do not need to have any experience of Linux to install Agua. It simply requires you to cut and paste a series of commands into the command line.

NAVIGATION BAR

|  |
| --- |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**1. AMI Launch Steps**](http://www.aguadev.org/confluence/display/getit/1.+AMI+Launch+Steps) |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**1.1. Create EC2 account**](http://www.aguadev.org/confluence/display/getit/1.1.+Create+EC2+account) |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**1.2. Generate Credentials**](http://www.aguadev.org/confluence/display/getit/1.2.+Generate+Credentials) |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**1.3. Launch Agua AMI**](http://www.aguadev.org/confluence/display/getit/1.3.+Launch+Agua+AMI) |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**2. Installation Steps**](http://www.aguadev.org/confluence/display/getit/2.+Installation+Steps) |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**2.1. System requirements**](http://www.aguadev.org/confluence/display/getit/2.1.+System+requirements) |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**2.2. Install Perl modules**](http://www.aguadev.org/confluence/display/getit/2.2.+Install+Perl+modules) |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**2.3. Download Agua**](http://www.aguadev.org/confluence/display/getit/2.3.+Download+Agua) |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | [**2.4. Install Agua**](http://www.aguadev.org/confluence/display/getit/2.4.+Install+Agua) |

# **1. AMI Launch Steps**

### **Running Agua using the EC2 AMI**

AMI stands for Amazon Machine Image, which is a frozen copy of a server's operating system and all the data on its hard disks. Agua's installation procedure has been tested on Ubuntu versions 10.04 and 11.04.

Other varieties of Linux may be supported in the future. In the meantime, for a trouble-free installation, it is advisable to use the currently supported varieties.

You will need a credit card and an internet connect for the next two steps.

#### 1. [Create an Amazon EC2 account](http://www.aguadev.org/confluence/display/getit/1.1.+Create+EC2+account)

#### 2. [Generate Amazon EC2 credentials](http://www.aguadev.org/confluence/display/getit/1.2.+Generate+Credentials)

#### 3. [Launch the Agua AMI](http://www.aguadev.org/confluence/display/getit/1.3.+Launch+Agua+AMI).