g-INFO portal

Doan Trung Tung

Aurélien BERNARD, Ana Lucia DA-COSTA, Vincent BLOCH, Thanh-Hoa LE, Yannick LEGRE, Lydia MAIGNE, Jean SALZEMANN, Hong-Quang NGUYEN, Vincent BRETON

Outline

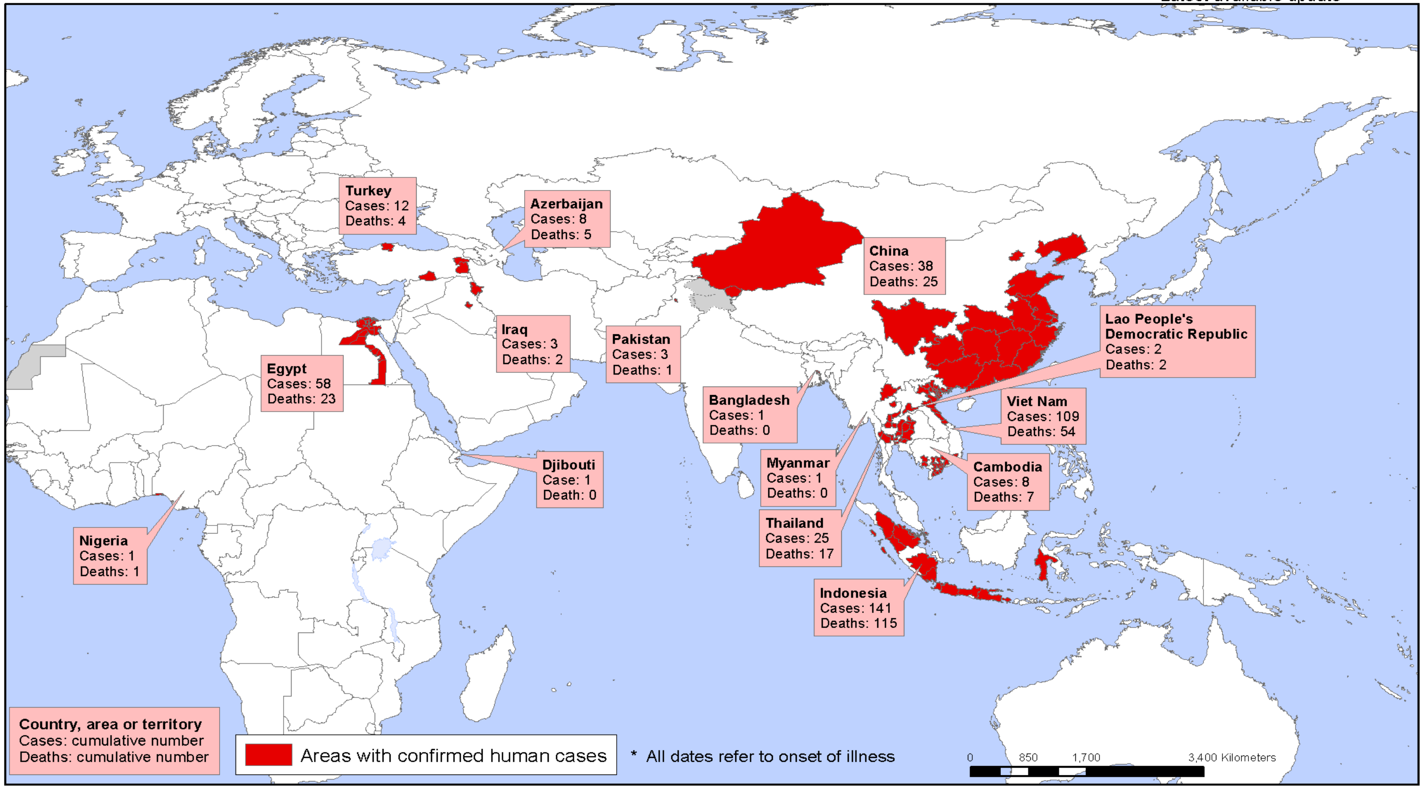
- Introduction
- Overview of g-INFO
- Implementation of g-INFO
- Conclusions and perspectives

Why g-INFO?

- H5N1 (avian flu)

Areas with confirmed human cases of H5N1 avian influenza since 2003 *

Status as of 11 March 2009
Latest available update



* All dates refer to onset of illness

World Health Organization

The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement. © WHO 2009. All rights reserved

Data Source: WHO
Map Production: Public Health Information and Geographic Information System (GIS)
World Health Organization

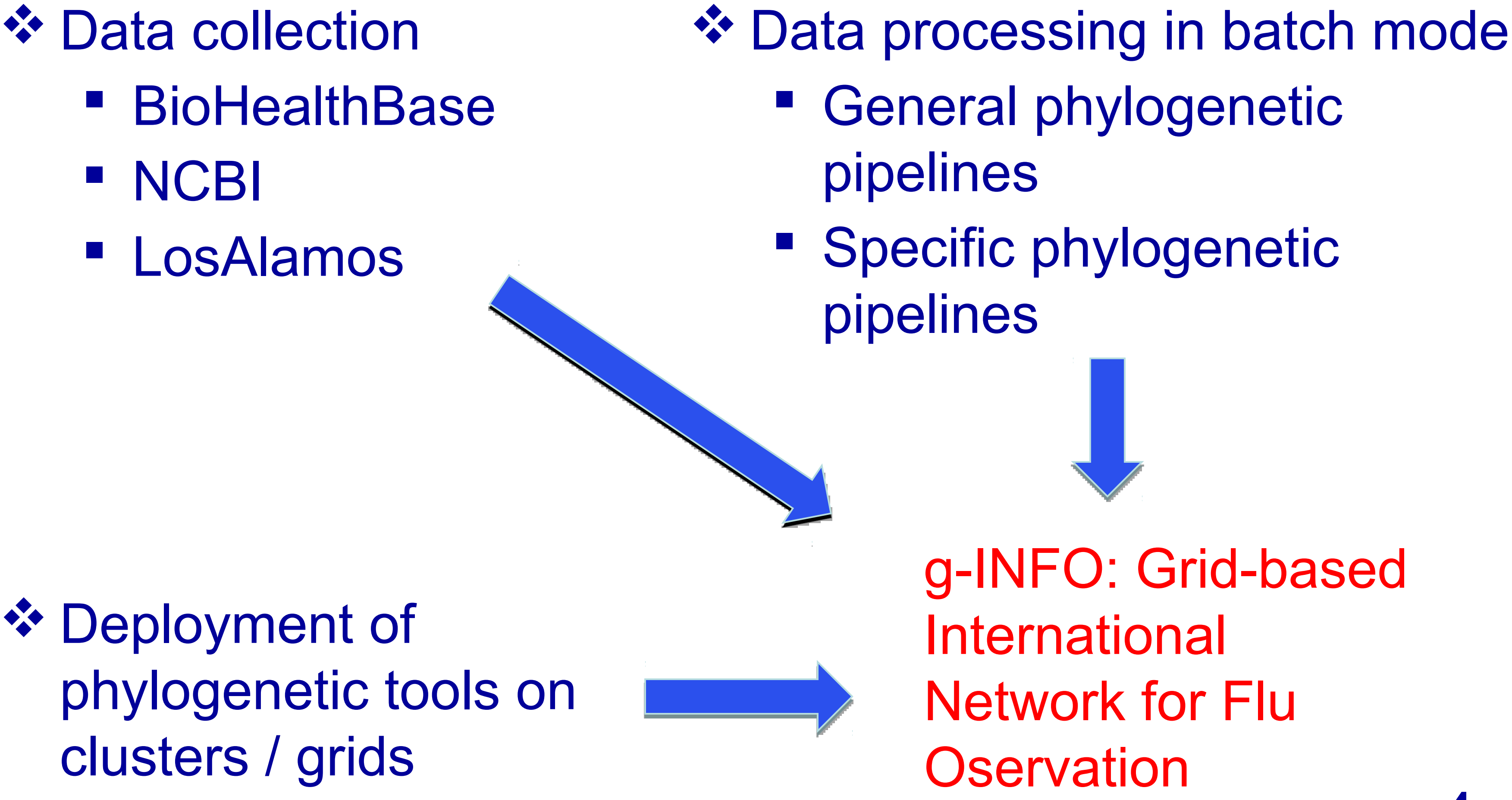
262 deaths
436 cases

WHO - July 2009

287 deaths
486 cases

WHO – March 2010

Influenza surveillance



- Data collection
  - BioHealthBase
  - NCBI
  - LosAlamos

- Data processing in batch mode
  - General phylogenetic pipelines
  - Specific phylogenetic pipelines

- Deployment of phylogenetic tools on clusters / grids

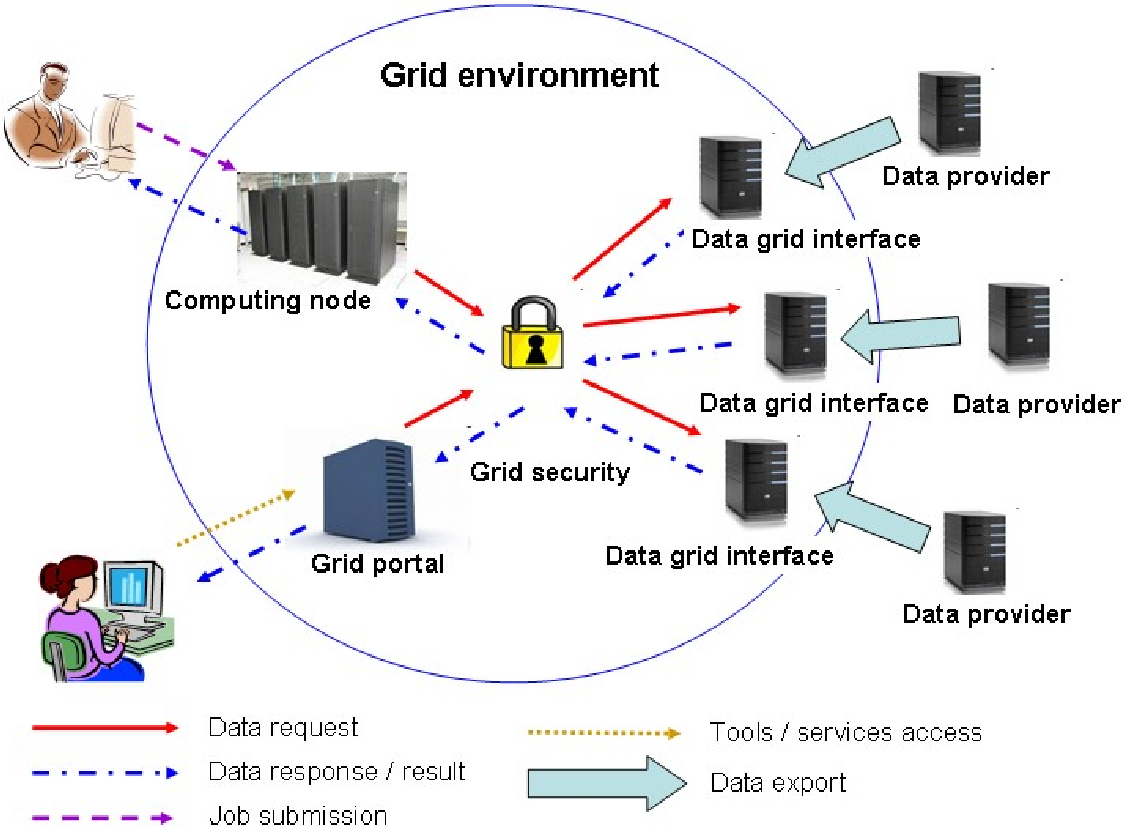
g-INFO: Grid-based International Network for Flu Oservation

g-INFO’s overview

g-INFO’s goals

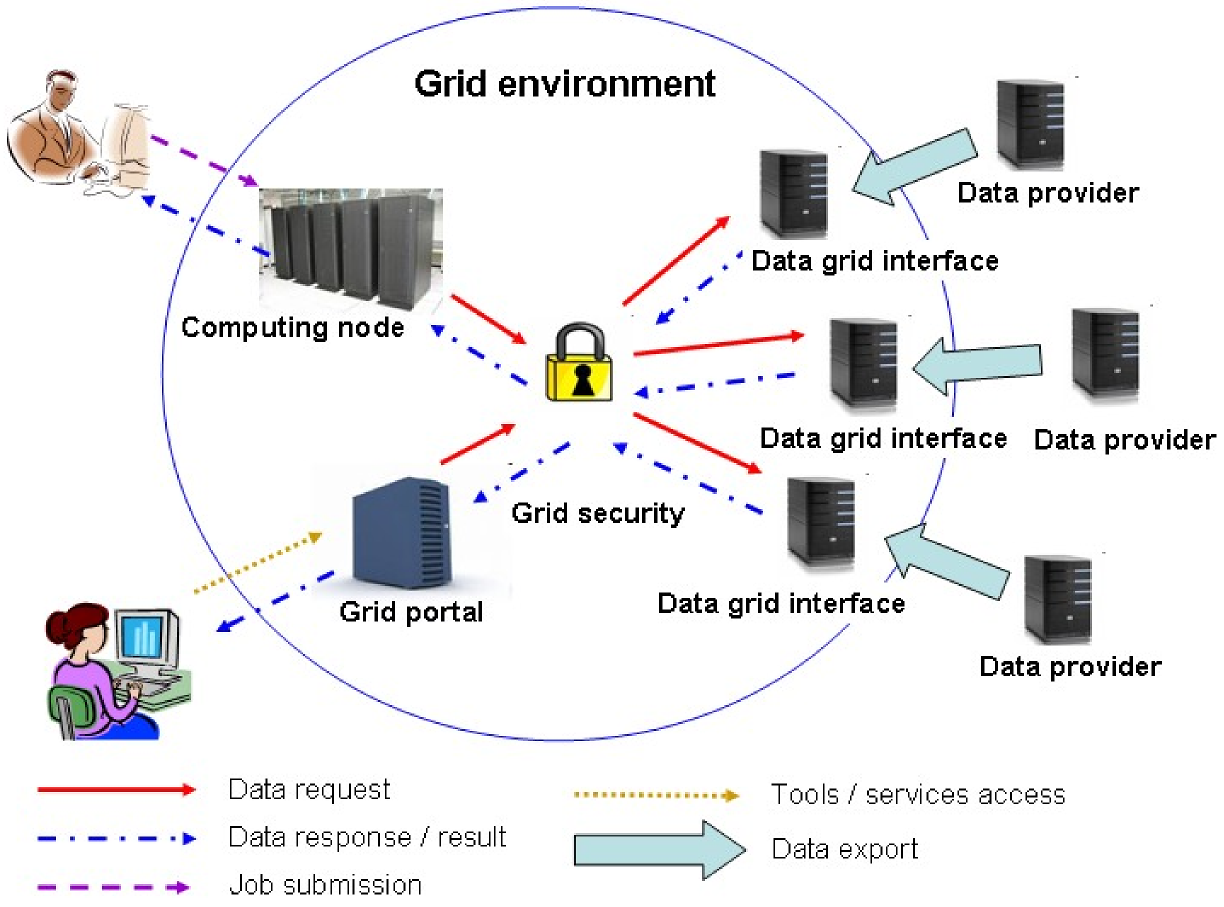
- Integration of influenza virus data sources into a federation of databases
- Automatic phylogenetic pipelines
- Specific molecular epidemiology studies

Architecture of g-INFO system



- Each data provider has its own server(s) to store his data
- Data provider export only selected data to a data grid interface server
- The data exported is integrated in a common schema on the interface servers
- Providers can keep the privilege of granting access rights to their data

Architecture of g-INFO system



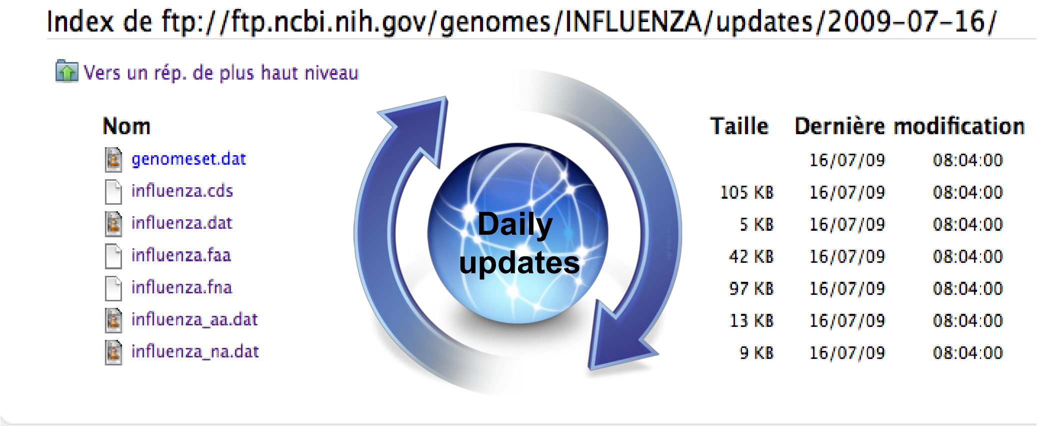
- Epidemiologic pipelines will be deployed on the grid
  - BLAST
  - Alignment
  - Phylogenetic trees
  - Visualisation
  - ... and more

g-INFO’s implementation

Data collection

```
>ABV25634
MKAILLVLLCAFAATNADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCRLGGIAPLQLG
KCNIAGWLLGNPECDLLLTVSSWSYIVETSNSDNGTCYPGDFIDYEELREQLSSVSSFEKFEIFPKTSSW
PNHETTRGVTAACPYAGASSFYRNLLWLVKKENSYPKLSKSYVNNKGKEVLVLWGVHHPPTSTDQQSLYQ
NADAYVSVGSSKYDRRFTPEIAARPKVRGQAGRMNYYWTLLEPGDTITFEATGNLVAPRYAFALNRGSES
GIITSDAPVHDCDTKCQTPHGAINSSLPFQNIHPVTIGECPKYVKSTKLRMVTGLRNIPSIQSRGLFGAI
AGFIEGGWTGLIDGWYGYHHQNGQGSGYAADQKSTQNAIDGITNKVNSVIEKMNTQFTVVGKEFNNLERR
IKNLNKKVDDGFLDVWTYNAELLVLLENERTLDFHDSNVKNLYEKARSQLRNNAKEIGNGCFEFYHKCDD
ACMESVRNGTYDYPKYSEESKLNREEIDGVKLESMMVYQILAIYSTVASSLVLLVSLGAISFWMCSNGSL
QCRICI
```

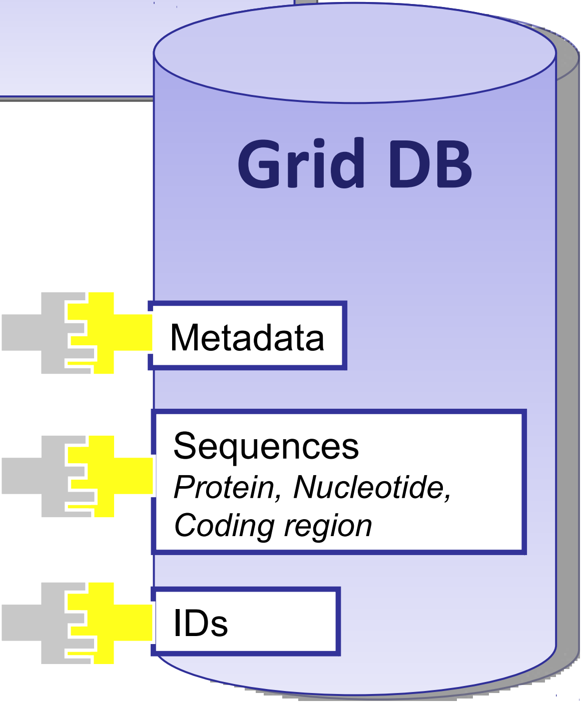
FTP NCBI



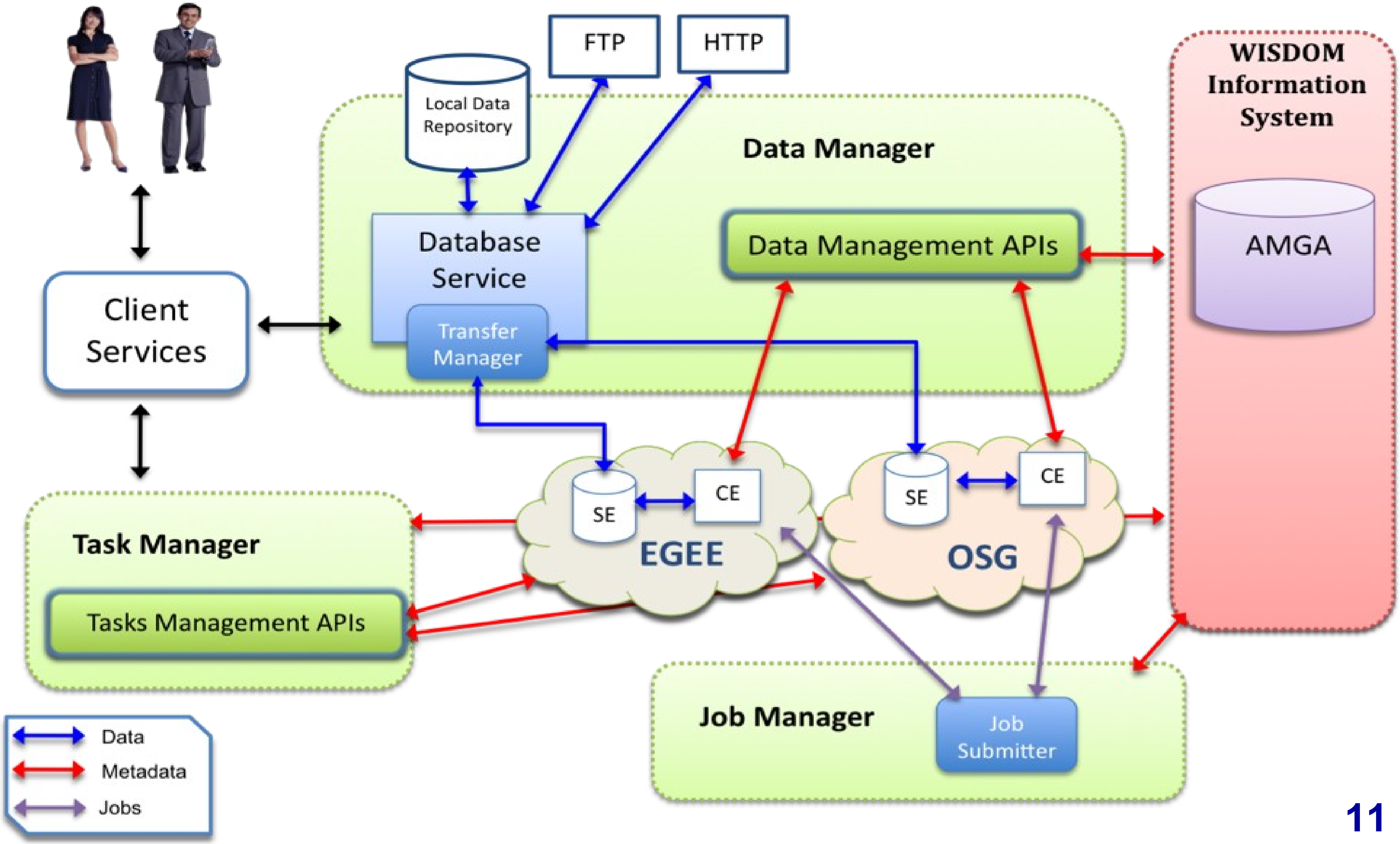
Index de ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/updates/2009-07-16/

Vers un rép. de plus haut niveau

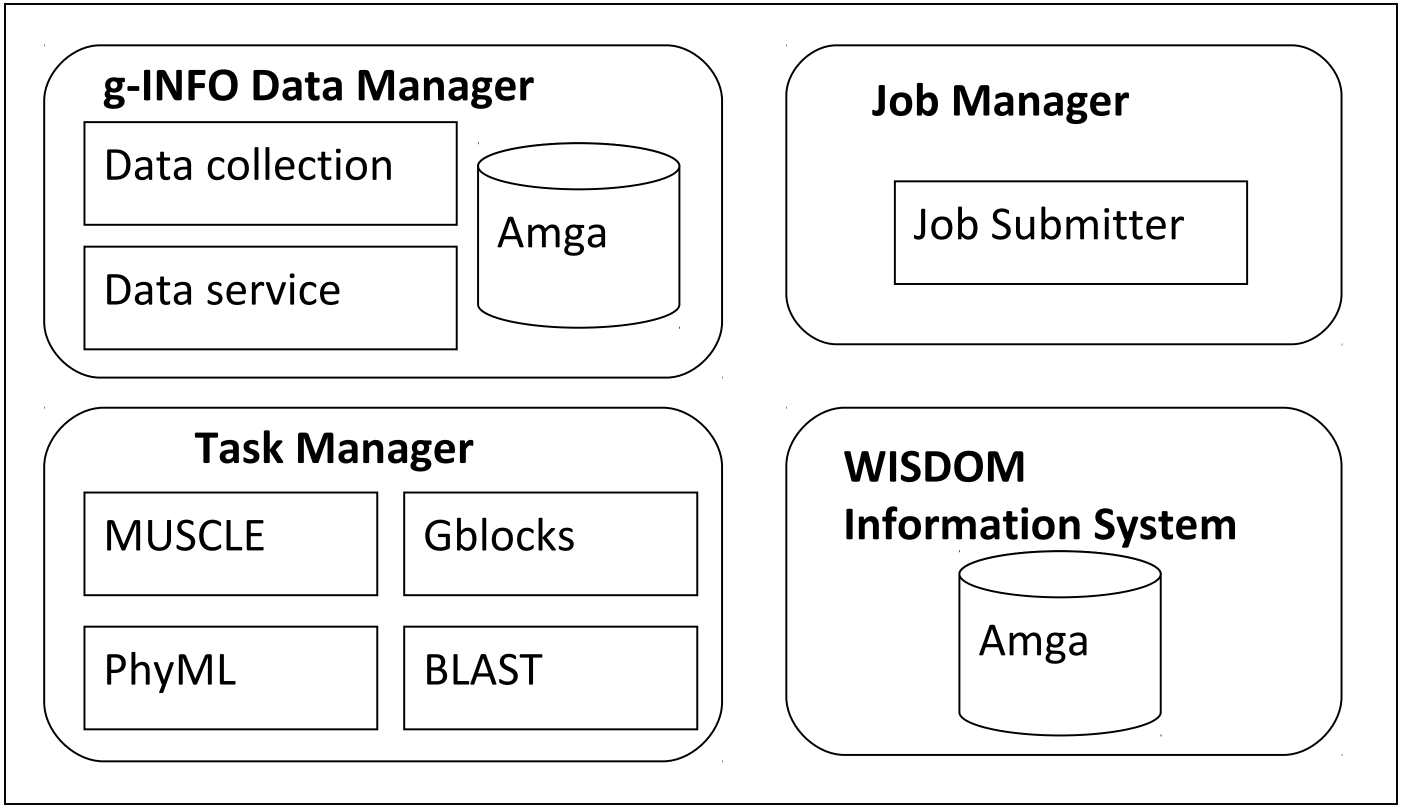
Nom	Taille	Dernière modification	
genomeset.dat		16/07/09	08:04:00
influenza.cds	105 KB	16/07/09	08:04:00
influenza.dat	5 KB	16/07/09	08:04:00
influenza.faa	42 KB	16/07/09	08:04:00
influenza.fna	97 KB	16/07/09	08:04:00
influenza_aa.dat	13 KB	16/07/09	08:04:00
influenza_na.dat	9 KB	16/07/09	08:04:00



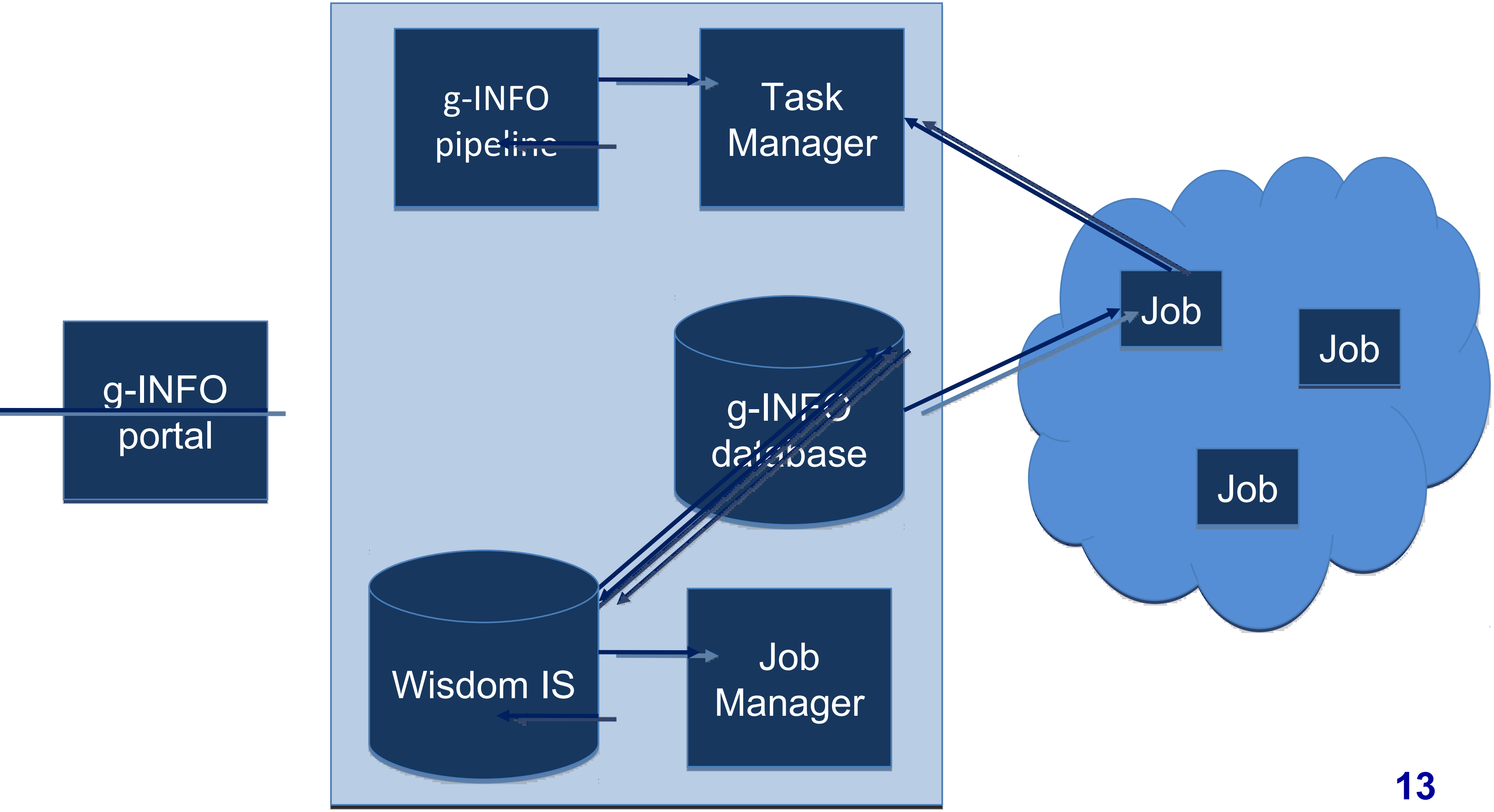
WISDOM Production Environment



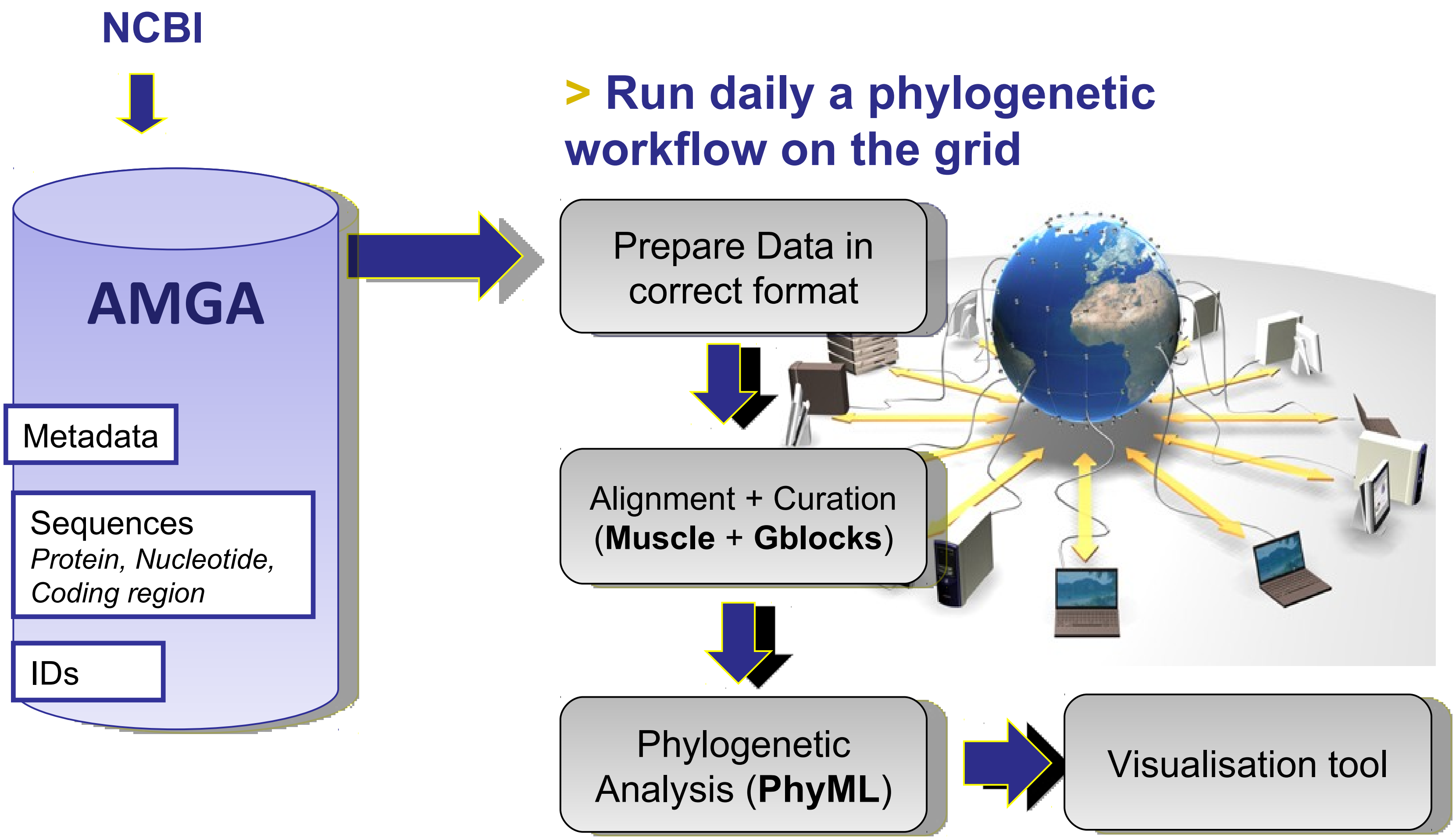
Integration of g-INFO into WPE



Automatic phylogenetic pipeline



Automatic phylogenetic pipeline



NCBI

AMGA

Metadata

Sequences
Protein, Nucleotide,
Coding region

IDs

> Run daily a phylogenetic workflow on the grid

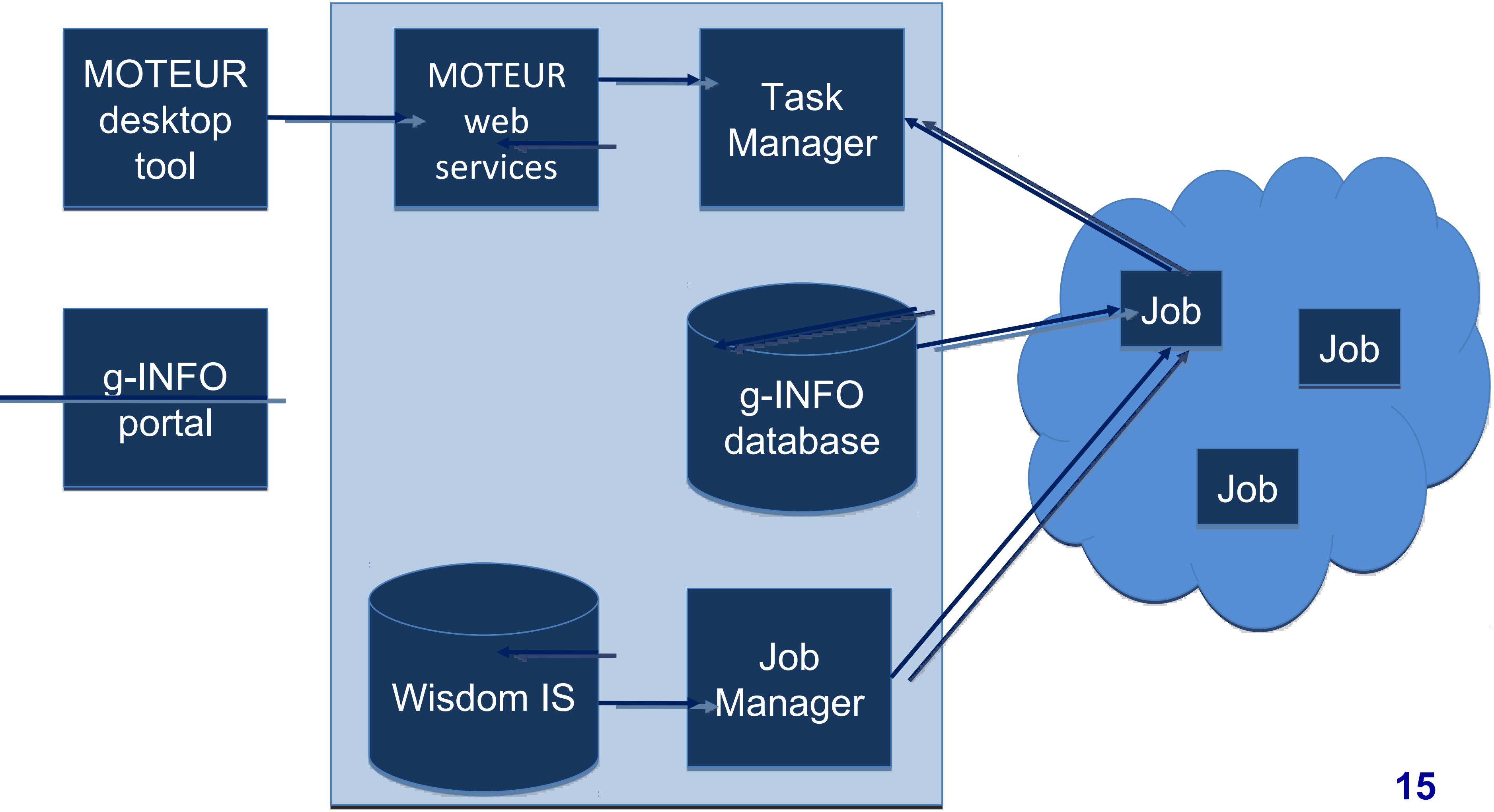
Prepare Data in correct format

Alignment + Curation (**Muscle + Gblocks**)

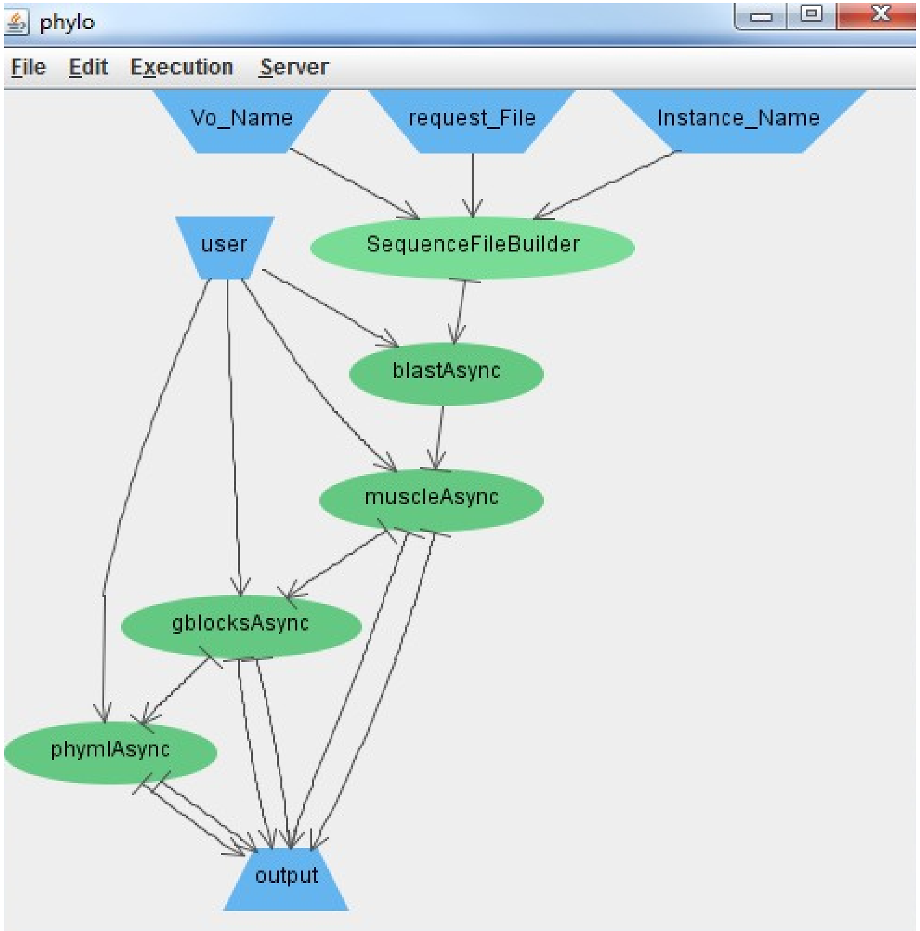
Phylogenetic Analysis (**PhyML**)

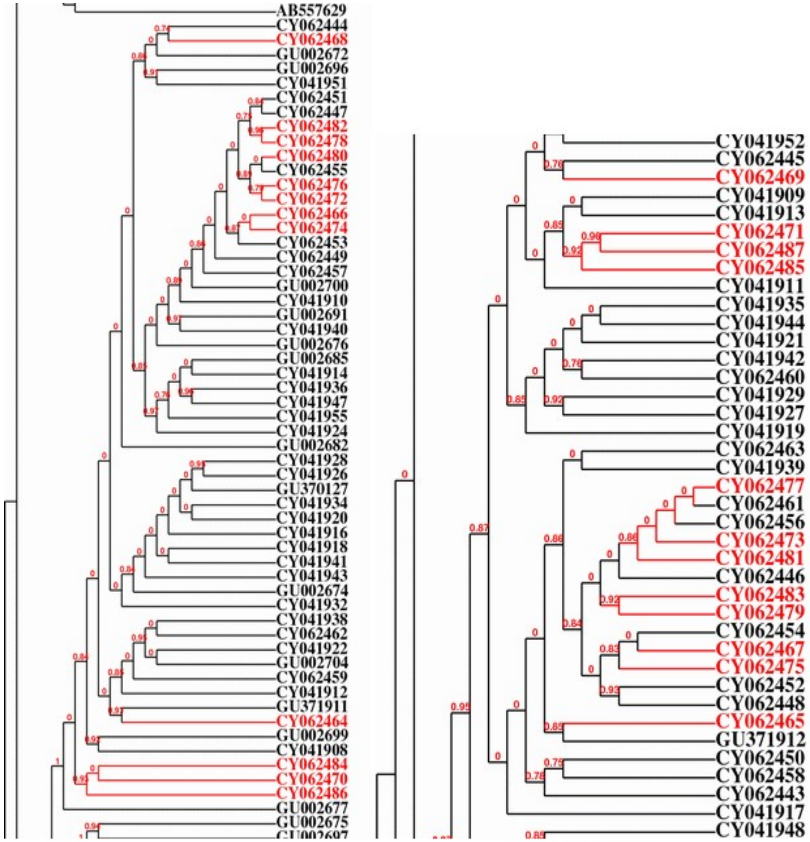
Visualisation tool

Manual phylogenetic workflow



Workflow execution example





HA - segment

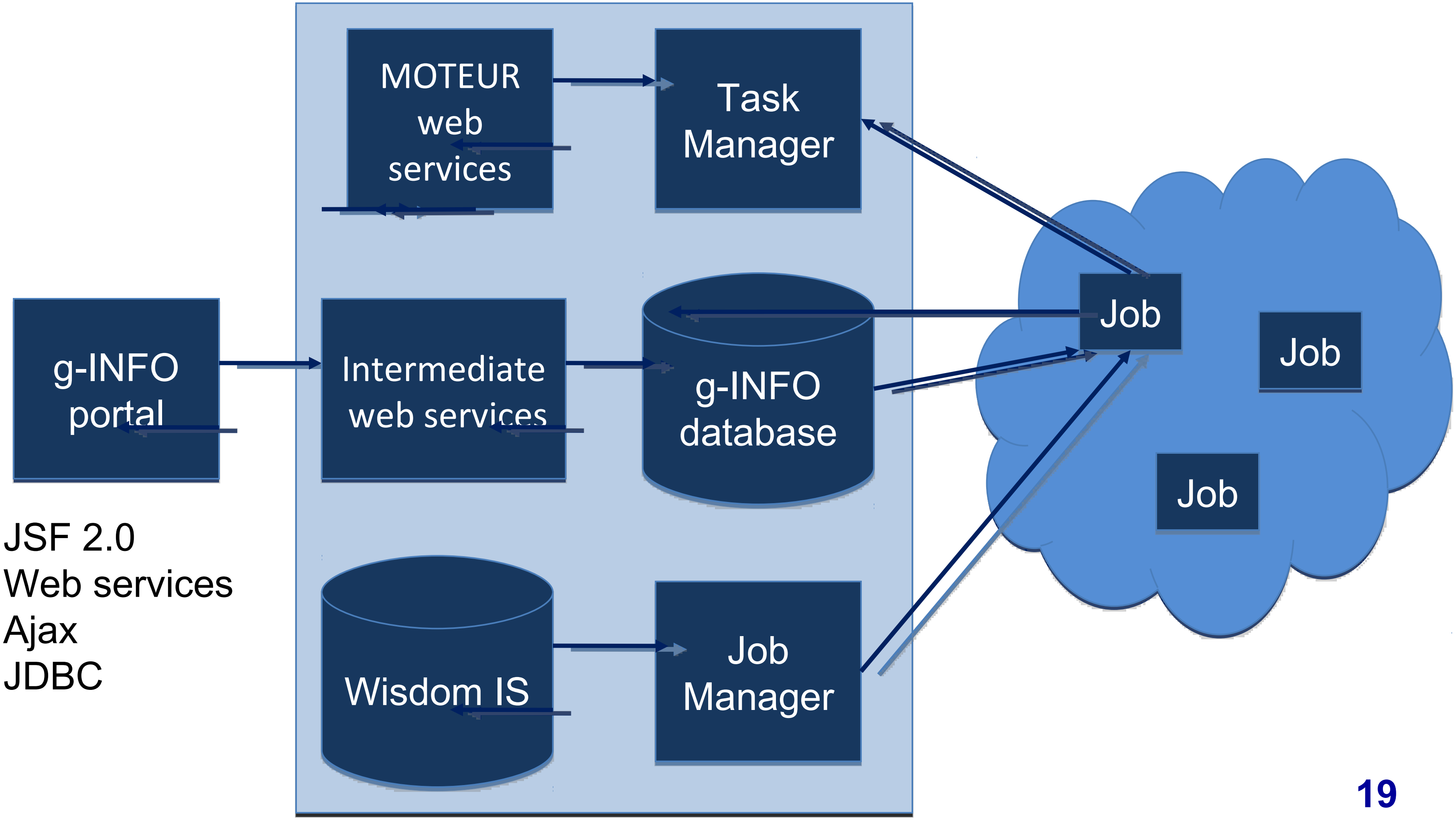
NA - segment

g-INFO portal

g-INFO portal

- Collaboration among IFI, IOIT and HPC:
  - IFI: web services to interact between the portal and the system
  - IOIT: design and develop the portal
  - HPC: visualization tool
- Technologies:
  - JSF 2.0, Ajax, web services, Java aplet, …

g-INFO portal



JSF 2.0
Web services
Ajax
JDBC

g-INFO portal – home page

g-INFO: Home

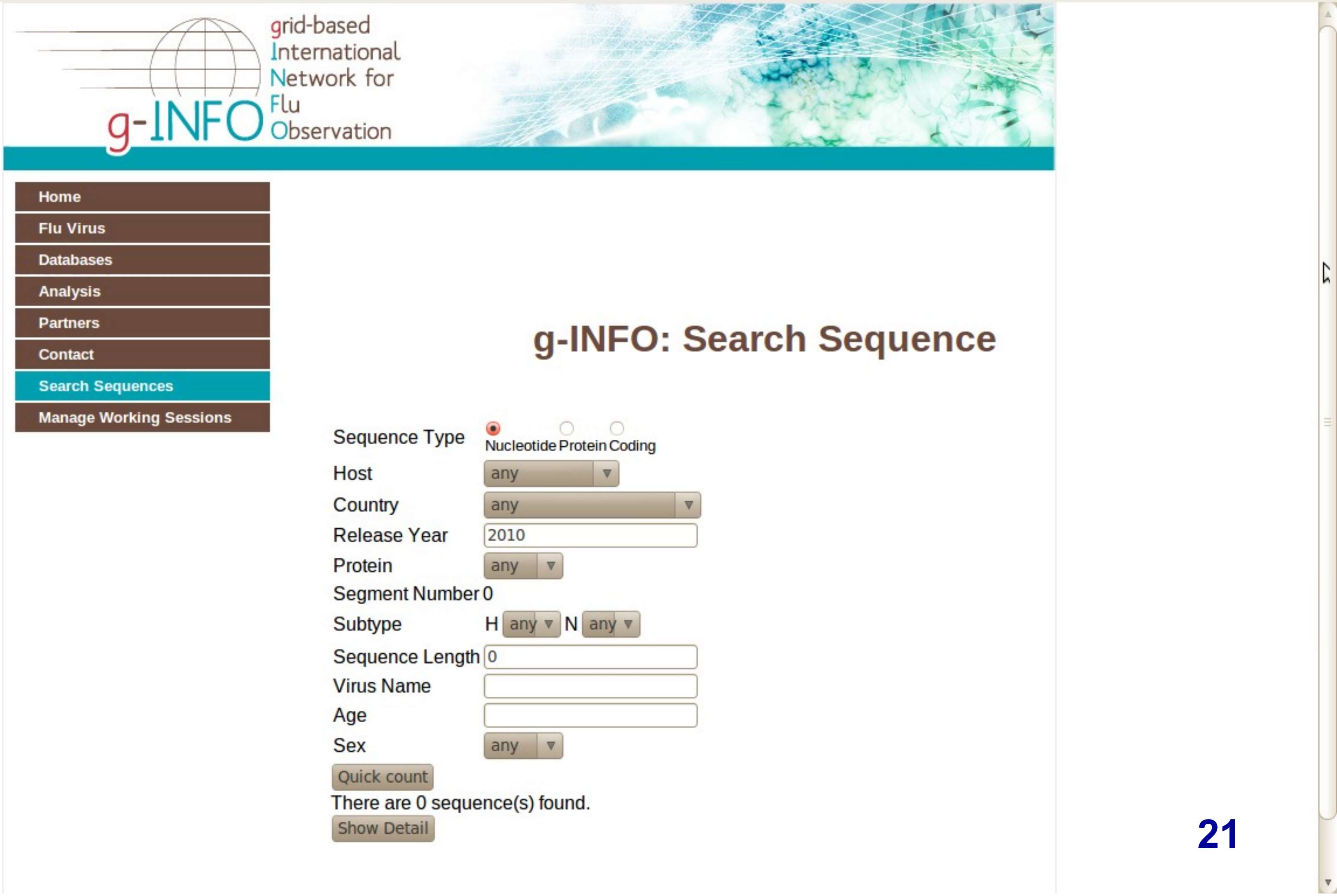
Recent years have seen the emergence of diseases which have spread very quickly around the world, either through human travel, like SARS and SIV(H1N1), or animal migration, like avian flu (H5N1) or more recently, the swine flu outbreak that has been classified as a "pandemic" by WHO in response to its world-wide geographic Spread.

Among the biggest challenges from emerging infectious diseases, is the relation to early detection and surveillance of the diseases, as new cases can appear anywhere. This is due to the globalization of exchanges and the circulation of people and animals around the world, as recently demonstrated by the avian flu epidemics. An international collaboration of research teams in Europe and Asia has been exploring some innovative in silico approaches to better tackle flu, taking advantage of the very large computing resources available on international Grid infrastructures. Based on current H1N1 pandemic example, it is expected to have an impact by adding a new weapon to researchers' arsenal: the grid.

Existing data sources have been integrated towards a global surveillance network for molecular epidemiology, based on Service Oriented Architecture (SOA) and Grid technologies. The idea is to dynamically analyze the molecular biology data, made available on public databases using computing, storage and automatic updating services offered by grid technology.

Design HealthGrid | Valid XHTML | Valid CSS | CC-BY-SA

g-INFO portal - search

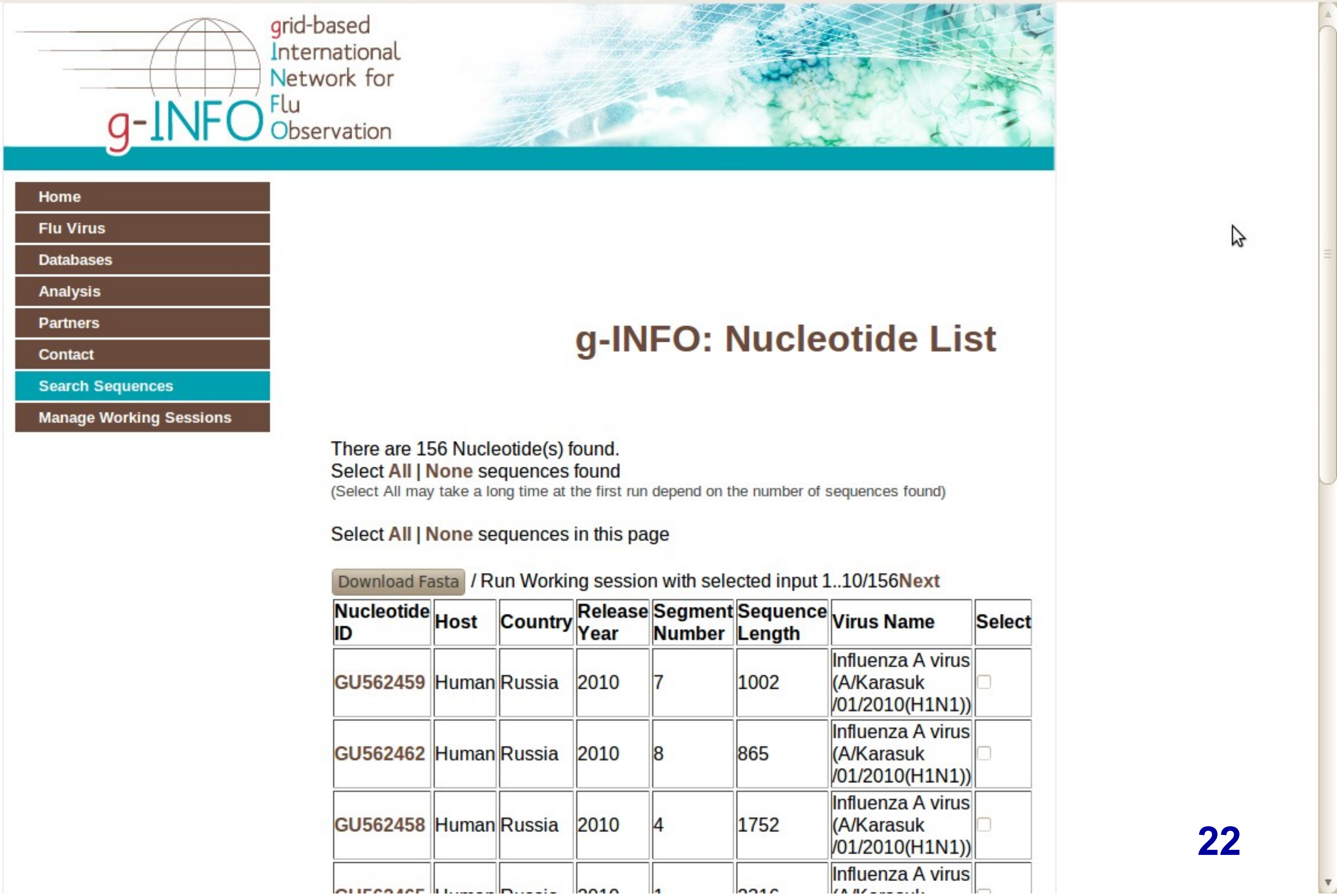


g-INFO: Search Sequence

Sequence Type ● Nucleotide ○ Protein ○ Coding
Host: any
Country: any
Release Year: 2010
Protein: any
Segment Number 0
Subtype: H any N any
Sequence Length: 0
Virus Name:
Age:
Sex: any

There are 0 sequence(s) found.

g-INFO portal – search results



g-INFO: Nucleotide List

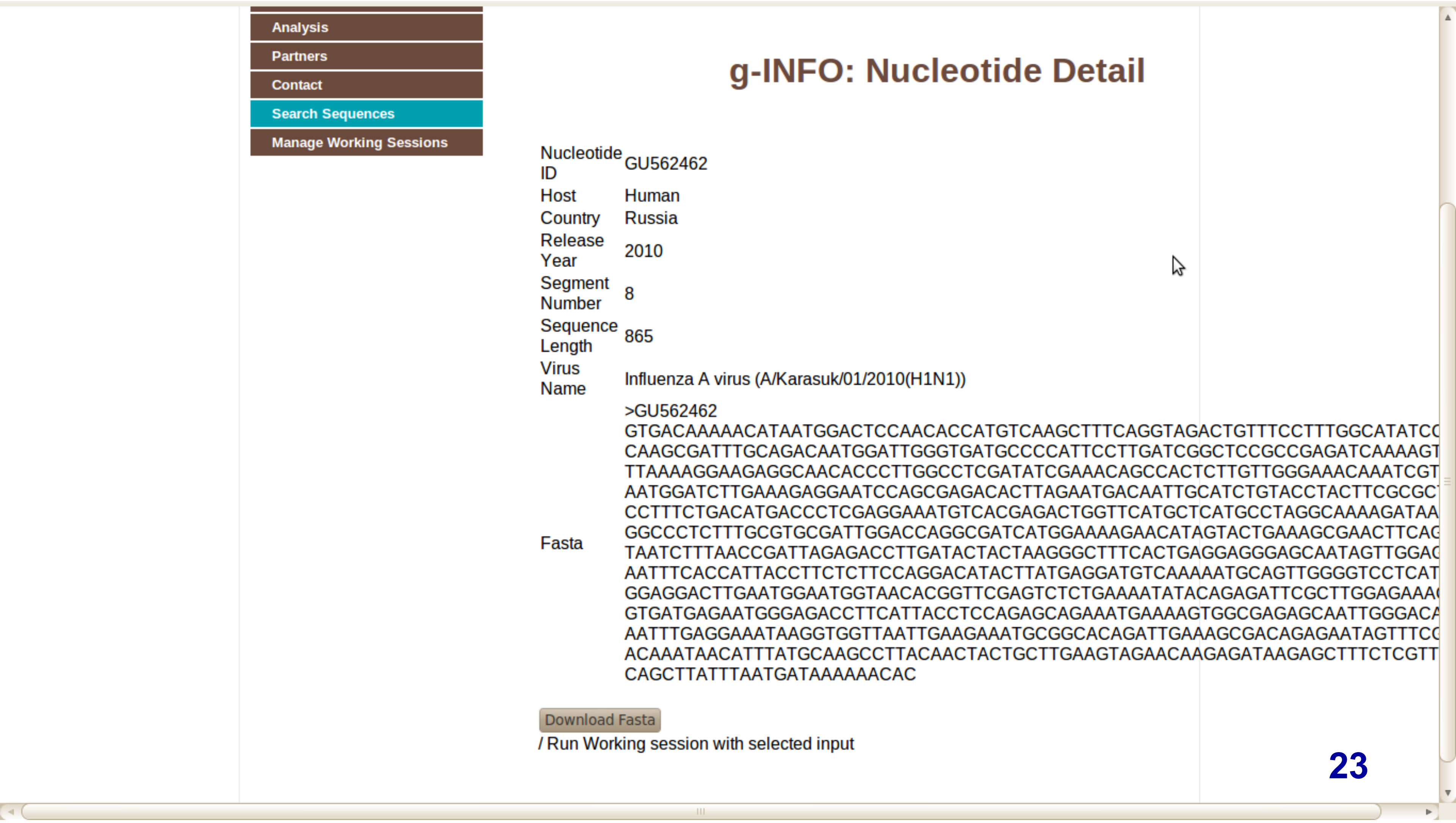
There are 156 Nucleotide(s) found.
Select **All** | **None** sequences found
(Select All may take a long time at the first run depend on the number of sequences found)

Select **All** | **None** sequences in this page

Download Fasta / Run Working session with selected input 1..10/156**Next**

Nucleotide ID	Host	Country	Release Year	Segment Number	Sequence Length	Virus Name	Select
GU562459	Human	Russia	2010	7	1002	Influenza A virus (A/Karasuk /01/2010(H1N1))	☐
GU562462	Human	Russia	2010	8	865	Influenza A virus (A/Karasuk /01/2010(H1N1))	☐
GU562458	Human	Russia	2010	4	1752	Influenza A virus (A/Karasuk /01/2010(H1N1))	☐
						Influenza A virus	

g-INFO portal – search results



g-INFO: Nucleotide Detail

Nucleotide ID	GU562462
Host	Human
Country	Russia
Release Year	2010
Segment Number	8
Sequence Length	865
Virus Name	Influenza A virus (A/Karasuk/01/2010(H1N1))
Fasta	>GU562462 GTGACAAAAACATAATGGACTCCAACACCATGTCAAGCTTTCAGGTAGACTGTTTCCTTTGGCATATC CAAGCGATTTGCAGACAATGGATTGGGTGATGCCCCATTCCTTGATCGGCTCCGCCGAGATCAAAAGT TTAAAAGGAAGAGGCAACACCCTTGGCCTCGATATCGAAACAGCCACTCTTGTTGGGAAACAAATCGT AATGGATCTTGAAAGAGGAATCCAGCGAGACACTTAGAATGACAATTGCATCTGTACCTACTTCGCGC CCTTTCTGACATGACCCTCGAGGAAATGTCACGAGACTGGTTCATGCTCATGCCTAGGCAAAAGATAA GGCCCTCTTTGCGTGCGATTGGACCAGGCGATCATGGAAAAGAACATAGTACTGAAAGCGAACTTCAG TAATCTTTAACCGATTAGAGACCTTGATACTACTAAGGGCTTTCACTGAGGAGGGAGCAATAGTTGGA AATTTCACCATTACCTTCTCTTCCAGGACATACTTATGAGGATGTCAAAAATGCAGTTGGGGTCCTCAT GGAGGACTTGAATGGAATGGTAACACGGTTCGAGTCTCTGAAAATATACAGAGATTCGCTTGGAGAAA GTGATGAGAATGGGAGACCTTCATTACCTCCAGAGCAGAAATGAAAAGTGGCGAGAGCAATTGGGACA AATTTGAGGAAATAAGGTGGTTAATTGAAGAAATGCGGCACAGATTGAAAGCGACAGAGAATAGTTTC ACAAATAACATTTATGCAAGCCTTACAACTACTGCTTGAAGTAGAACAAGAGATAAGAGCTTTCTCGTT CAGCTTATTTAATGATAAAAAACAC

/ Run Working session with selected input

g-INFO portal – working sessions

g-INFO: Working Sessions

This is your current working session list:

Working Session ID	Status	Start Date	End Date
WS03			
WorkingSession03			
WS04			
WorkingSessionAbc			
WorkingSessionGogogo			
hieu01			
WorkingSession_19Oct			
WorkingSession 29Oct			
WorkingSession01			
WorkingSession04	running	2010-11-10 11:10:42	
November 11	running	2010-11-11 10:19:41	
phien lam viec	running	2010-11-11 11:03:14	
nov 15	running	2010-11-15 03:05:04	
TestWS	running	2010-11-15 10:53:18	
abc	running	2010-11-16 08:59:12	
WorkingSession05	running	2010-11-26 11:20:03	

g-INFO portal – define working session template

g-INFO: Create New WorkFlow Template

WorkFlow Template Name: sample wft

Design HealthGrid | Valid XHTML | Valid CSS | CC-BY-SA

g-INFO portal – define working session template

g-INFO: Select Tool

WorkFlow Template Name sample wft

Component1 blast

Tool1 blast

Argument1 -p blastn -b 20 -a 4 -e 0.001

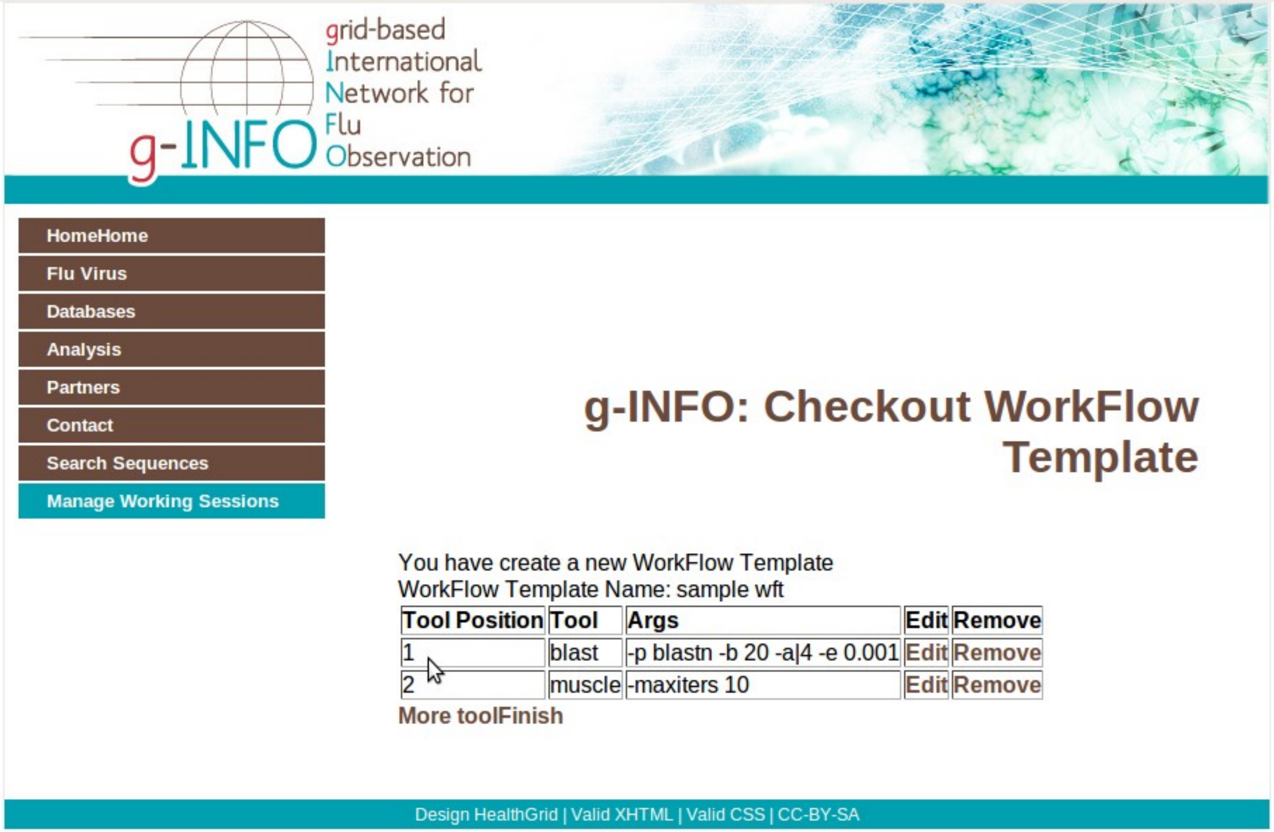
Save

g-INFO portal – define working session template

g-INFO: Select Tool

WorkFlow Template Name sample wft
Component2 alignment
Tool2 muscle
Argument2 -maxiters 10

g-INFO portal – define working session template

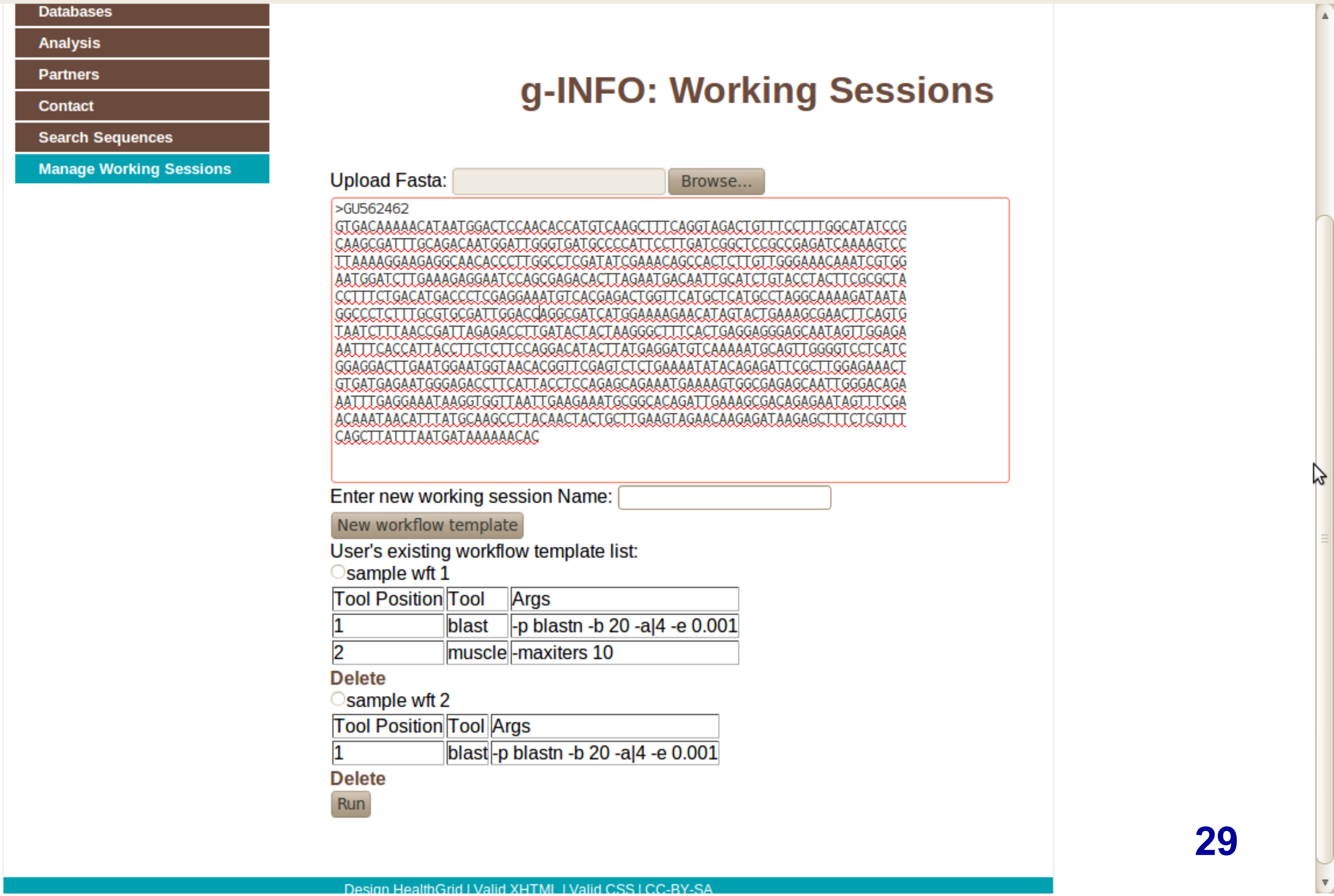


g-INFO: Checkout WorkFlow Template

You have create a new WorkFlow Template
WorkFlow Template Name: sample wft

Tool Position	Tool	Args	Edit	Remove
1	blast	-p blastn -b 20 -a 4 -e 0.001	Edit	Remove
2	muscle	-maxiters 10	Edit	Remove

g-INFO portal – run working session



g-INFO: Working Sessions

Upload Fasta:

```
>GU562462
GTGACAAAAACATAATGGACTCCAACACCATGTCAAGCTTTCAGGTAGACTGTTTCCTTTGGCATATCCG
CAAGCGATTTGCAGACAATGGATTGGGTGATGCCCCATTCCTTGATCGGCTCCGCCGAGATCAAAAGTCC
TTAAAAGGAAGAGGCAACACCCTTGGCCTCGATATCGAAACAGCCACTCTTGTTGGGAAACAAATCGTGG
AATGGATCTTGAAAGAGGAATCCAGCGAGACACTTAGAATGACAATTGCATCTGTACCTACTTCGCGCTA
CCTTTCTGACATGACCCTCGAGGAAATGTCACGAGACTGGTTCATGCTCATGCCTAGGCAAAAGATAATA
GGCCCTCTTTGCGTGCGATTGGACCAGGCGATCATGGAAAAGAACATAGTACTGAAAGCGAACTTCAGTG
TAATCTTTAACCGATTAGAGACCTTGATACTACTAAGGGCTTTCACTGAGGAGGGAGCAATAGTTGGAGA
AATTTCACCATTACCTTCTCTTCCAGGACATACTTATGAGGATGTCAAAAATGCAGTTGGGGTCCTCATC
GGAGGACTTGAATGGAATGGTAACACGGTTCGAGTCTCTGAAAATATACAGAGATTCGCTTGGAGAAACT
GTGATGAGAATGGGAGACCTTCATTACCTCCAGAGCAGAAATGAAAAGTGGCGAGAGCAATTGGGACAGA
AATTTGAGGAAATAAGGTGGTTAATTGAAGAAATGCGGCACAGATTGAAAGCGACAGAGAATAGTTTCGA
ACAAATAACATTTATGCAAGCCTTACAACTACTGCTTGAAGTAGAACAAGAGATAAGAGCTTTCTCGTTT
CAGCTTATTTAATGATAAAAAACAC
```

Enter new working session Name:

User's existing workflow template list:

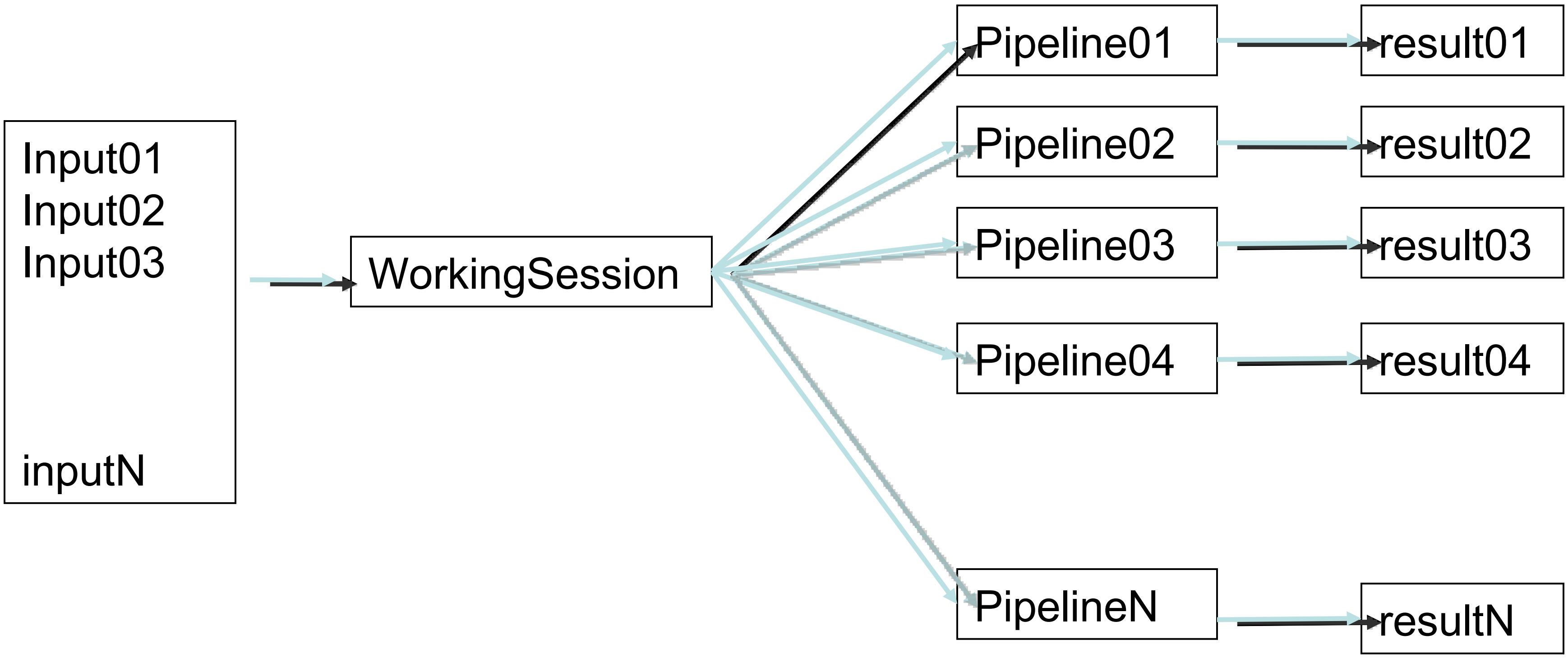
○ sample wft 1

Tool Position	Tool	Args
1	blast	-p blastn -b 20 -a|4 -e 0.001
2	muscle	-maxiters 10

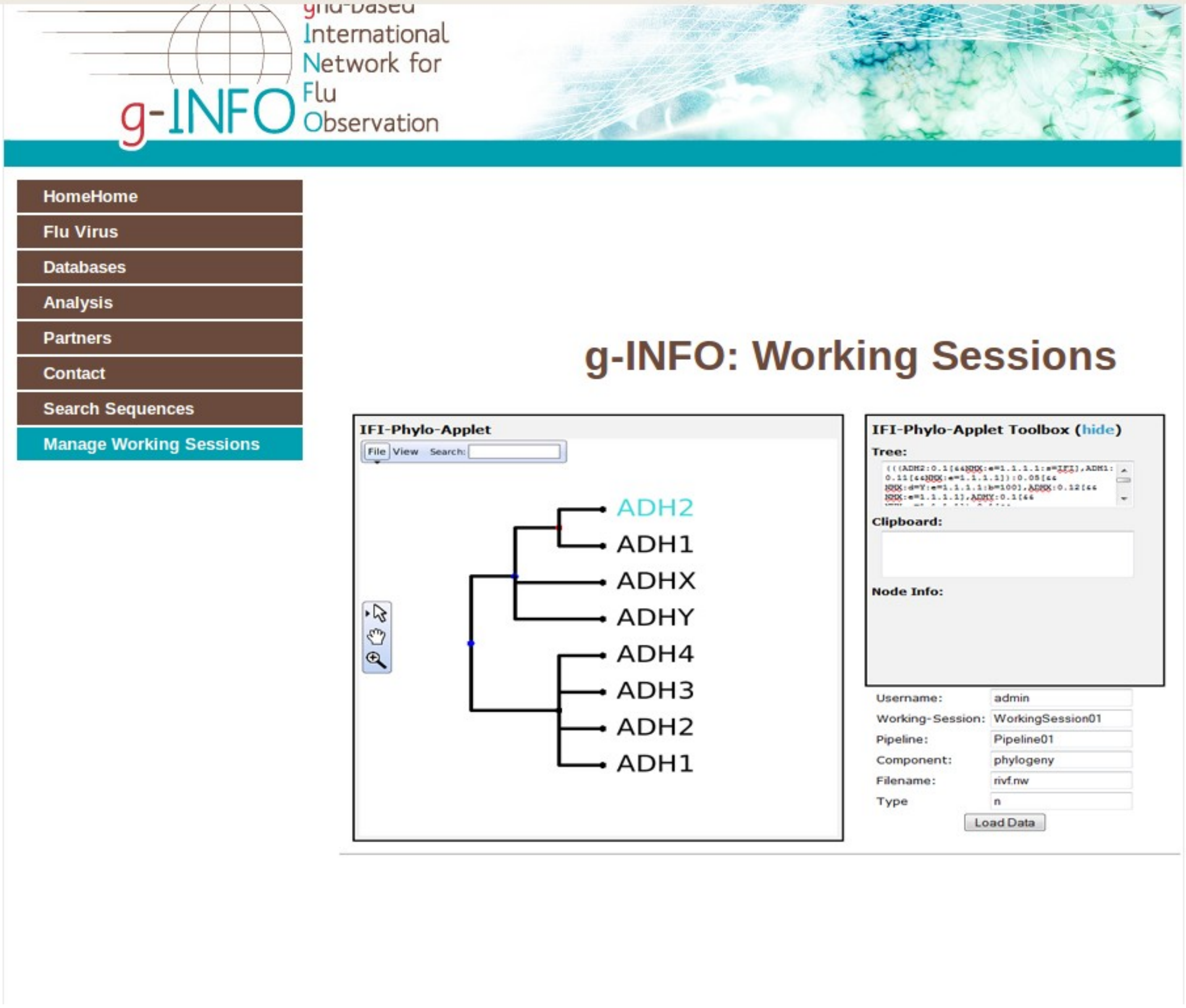
○ sample wft 2

Tool Position	Tool	Args
1	blast	-p blastn -b 20 -a|4 -e 0.001

g-INFO portal – run working session



g-INFO portal – visualization



g-INFO: Working Sessions

Tree:

```
(((ADH2:0.1[&&NHX:e=1.1.1.1:s=IFI],ADH1:
0.11[&&NHX:e=1.1.1.1]):0.05[&&
NHX:d=Y:e=1.1.1.1:b=100],ADHX:0.12[&&
NHX:e=1.1.1.1],ADHY:0.1[&&
```

Username:	admin
Working-Session:	WorkingSession01
Pipeline:	Pipeline01
Component:	phylogeny
Filename:	rivf.nw
Type	n

Conclusions

- A success in terms of international collaboration
- An example of developping grid application in Vietnam
- A complementary service for the public health research community

Perspectives

- Provide more tools and pipelines
- Import other database resources
- Improve system’s performance
- We are expecting the research community to contribute with more useful tools
- Can be applied for other emerging diseases

Thank you!