

# ACGRID 2009

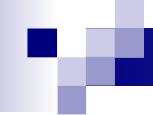
## Epidemiology on grid

Doan Trung Tung

Phd student – MSI-IFI & PCSV/UBP

# Content

- Introduction of AMGA
- Influenza A
- Global surveillance network
- NCBI database and synchronization from NCBI to AMGA
- Phylogenetic workflow
- Hands-on



# Introduction of AMGA

# Metadata on the GRID

- Metadata is data about data
- On the Grid: information about files
  - Describe files
  - Locate files based on their contents
- But also simplify DB access on the Grid
  - Many Grid applications need structured data
  - Many applications require only simple schemas
    - Can be modelled as metadata
- Main advantage: better integration with the Grid environment
  - Metadata Service is a Grid component
  - Grid security
  - Hide DB heterogeneity

# AMGA Implementation

- AMGA has been developed by ARDA
  - AMGA – ARDA Metadata Grid Application
- Now part of gLite middleware
  - Official Metadata Service for EGEE
  - First release with gLite 1.5
  - Also available as standalone component
- Expanding user community
  - HEP, Biomed, UNOSAT...

# Metadata concepts in AMGA (1)

- Metadata - List of attributes associated with entries
  - Metadata for virus sequences: host, subtype, year, ...
- Attribute – key/value pair with type information
  - Type – The type (int, float, string,...)
  - Name/Key – The name of the attribute
  - Value - Value of an entry's attribute

```
Query> getattr metadata/AB530996 host subtype year
>> AB530996
>> Avian
>> H5N1
>> 2009
```

# Metadata concepts in AMGA (2)

- Schema – A set of attributes
- Collection – A set of entries associated with a schema
- Think of schemas as tables, attributes as columns, entries as rows

```
Query> listattr metadata
>> host
>> varchar(30)
>> country
>> varchar(30)
>> year
>> varchar(20)
>> subtype
>> varchar(10)
```

```
Query> ls metadata
>> AB530989
>> AB530990
>> AB530991
>> AB530992
>> AB530993
>> AB530994
>> AB530995
>> AB530996
```

# AMGA features

- Dynamic schemas
  - Schemas can be modified at runtime by client
    - Create, delete schemas
    - Add, remove attributes
- Metadata organised as an hierarchy
  - Collections can contain sub-collections
  - Analogy to file system:
    - Collection  $\Leftrightarrow$  Directory; Entry  $\Leftrightarrow$  File
- Flexible queries
  - SQL-like query language
  - Joins between schemas

```
Query> ls test
>> /tung/users/test/metadata
>> /tung/users/test/nucleotide
>> /tung/users/test/protein
>> /tung/users/test/coding
>> /tung/users/test/updates
```

# Security

- Unix style permissions
- ACLs – Per-collection or per-entry.
- Secure connections – SSL
- Client Authentication based on
  - Username/password
  - General X509 certificates
  - Grid-proxy certificates
- Access control via a Virtual Organization Management System (VOMS)

# AMGA queries

## ■ Queries

- Query> find entry\_pattern  
query\_condition
- Query> selectattr column\_1\_query ...  
column\_n\_query query\_condition
- Query> updateattr attr\_1  
update\_query\_1 .... attr\_n  
update\_query\_n query\_condition
- Query> updateattr\_single attr\_1  
update\_query\_1 .... attr\_n  
update\_query\_n query\_condition

Type **help commands** for list of queries

# Query condition

- A query condition is a query which returns a boolean in order to select or not select an entry for retrieval or update. Examples are
  - /jobdir:events>1000 and  
/configdir:key=/jobdir:key like (/jobdir:FILE,  
"t%")
- Query conditions are used in the WHERE statements of the SQL queries which are passed to the backends.

# References to attributes

- References to attributes take the form:
  - <directory>:<attribute>
- Relative paths to the directory (which is synonyme for table or schema, here) are allowed. Examples:
  - Query> selectattr /test:t 'like(t, "Test%")'
  - Query> selectattr count(/test:t) 'like(t, "Test%")'

# Functions

- **lower(string):** Converts string to lower case.
- **upper(string):** Converts string to upper case.
- **count(x):** Aggregate function. Counts how often the attribute is set (not = NULL)
- **abs(x):** Absolute value of x.
- **sin(x):** The sine of x.
- **cos(x):** The cosine of x.
- **tan(x):** The tangens of x.
- **atan(x):** The arc-tangens of x.
- **sqrt(x):** The square root of x.
- **ln(x):** The natural logarithm of x.
- **log(x):** The base 10 logarithm of x.
- **rnd():** A random number between 0 and 1.
- **sum(x):** The aggregate sum of x.
- **max(x):** The aggregate maximum of x.
- **min(x):** The aggregate minimum of x.
- **avg(x):** The aggregate average of x.
- **length(string):** The length of the string.
- **pow(x, y):** x to the power of y.
- **mod(x, y):** x modulo y.
- **concat(str1, str2):** The concatenation of str1 and str2.
- **like(str, pattern):** Whether str is like pattern. The pattern is an SQL90 pattern.
- **substr(str, n, m):** The substring of length m of str starting at n.
- **isnull(arg):** Checks that the argument is NULL;
- **notnull(arg):** Checks that the argument is not NULL;
- **is(condition, t, e):** Evaluates to t if condition is fulfilled, otherwise to e.

# Joint tables

- The supported joins are left and right outer joins and the inner join. The following shows some examples
  - Query> selectattr /t1:num /t1:name /t2:num /t2:value '/t1:num = 1 join\_left\_on(/t1:, /t2:, /t1:num = /t2:num) limit 1'
  - Query> selectattr /t1:name /t2:value 'join\_right\_on(/t1:, /t2:, /t1:num = /t2:num) '

# Commands will be used

- cd
- ls
- pwd
- whoami
- listattr
- createdir
- rmdir
- rm
- addattr
- find
- selectattr
- getattr

# Conclusions

- AMGA – Metadata Service of gLite
  - Part of gLite 1.5
  - Useful for simplified DB access
  - Integrated on the Grid environment (Security)
- Replication/Federation under development
- Tests show good performance/scalability
- Already deployed by several Grid Applications
  - LHCb, ATLAS, Biomed, ...
  - DLibrary (next presentation)
- AMGA Web Site

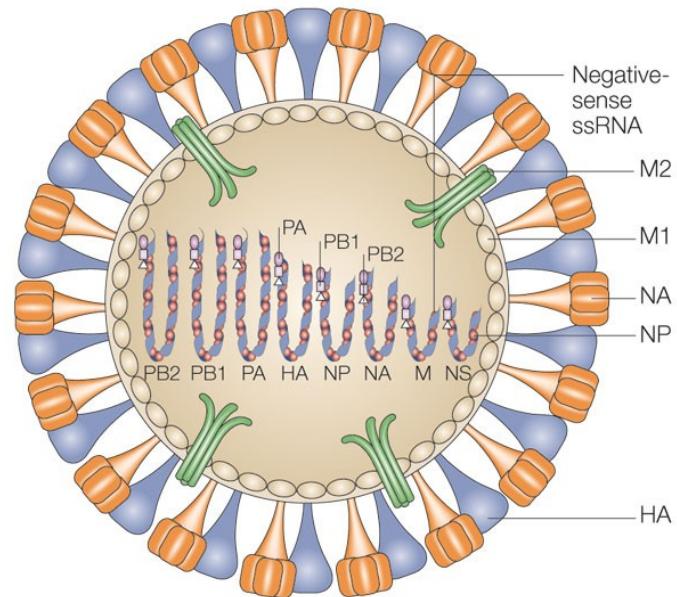
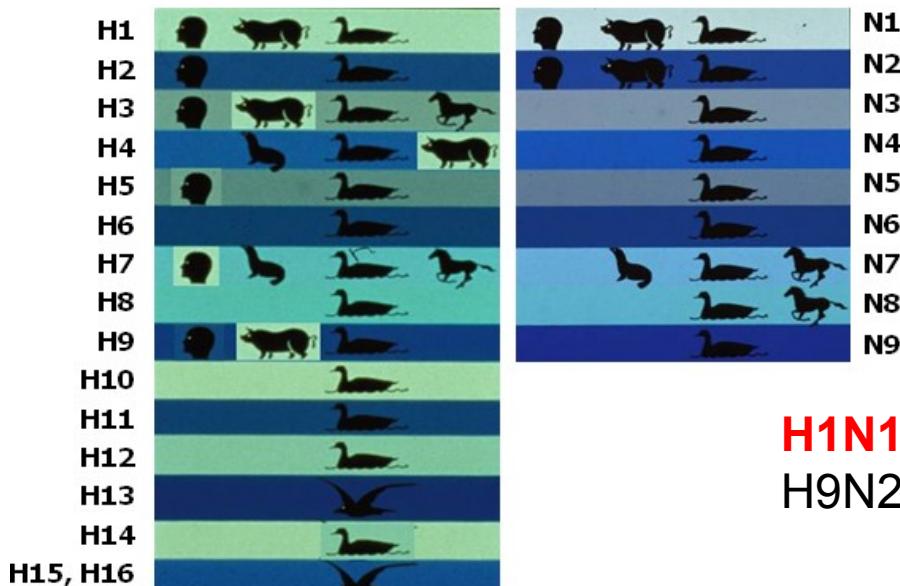
<http://project-arda-dev.web.cern.ch/project-arda-dev/metadata/>

# Context of Influenza A

# Current status of Influenza A

## ❖ Influenza A virus

- 8 nucleotides, 11 proteins
- 11 HA (hemagglutinin)
- 9 NA (neuraminidase)



Copyright © 2005 Nature Publishing Group  
Nature Reviews | Microbiology

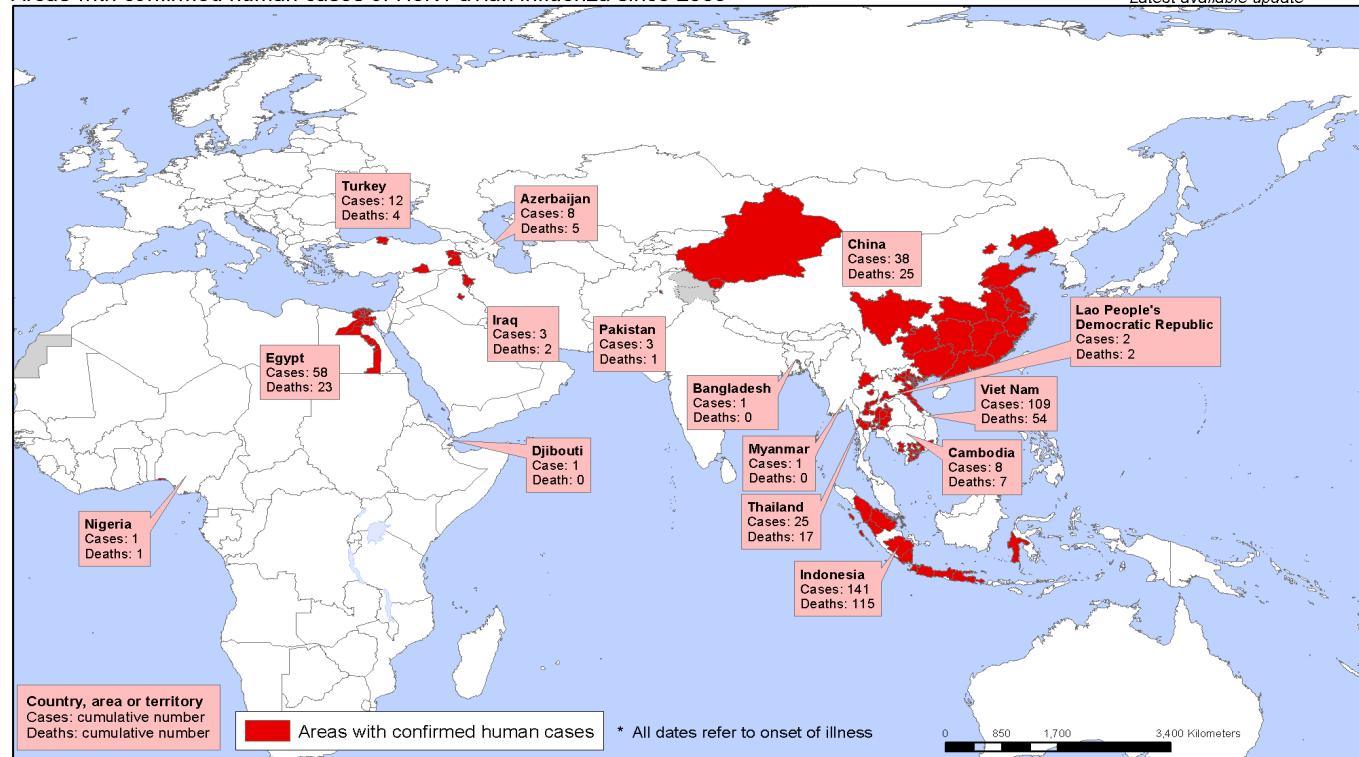
**H1N1**, H2N2, N3N2, **H5N1**, H7N7, H1N2,  
H9N2, H7N2, H7N3 and H10N7

# Current status of Influenza A

## ❖ H5N1 (avian influenza)

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

Status as of 11 March 2009  
Latest available update

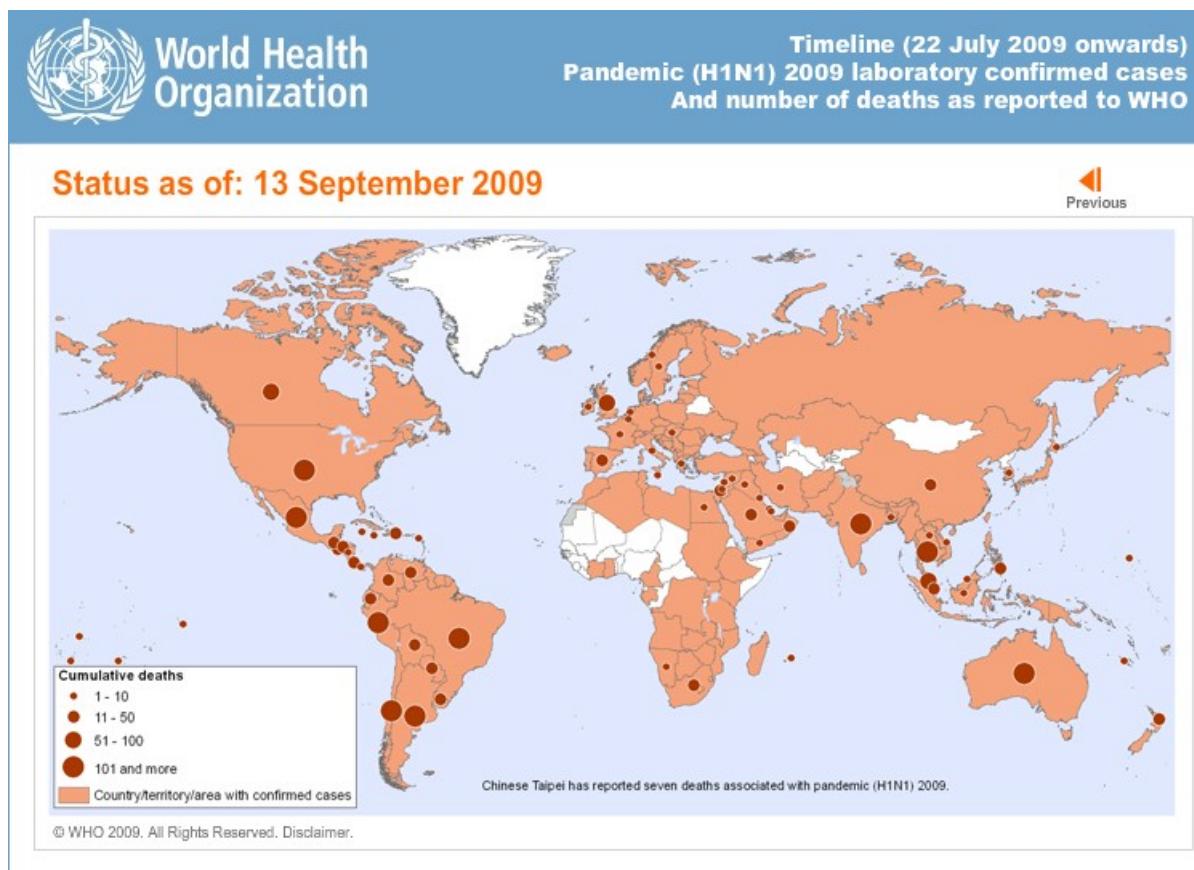


262 / 440

WHO - 31 August 2009

# Current status of Influenza A

## ❖ H1N1



382 / 89 921

WHO - 3 July 2009

3486 / 296471

WHO - 13 Sept 2009

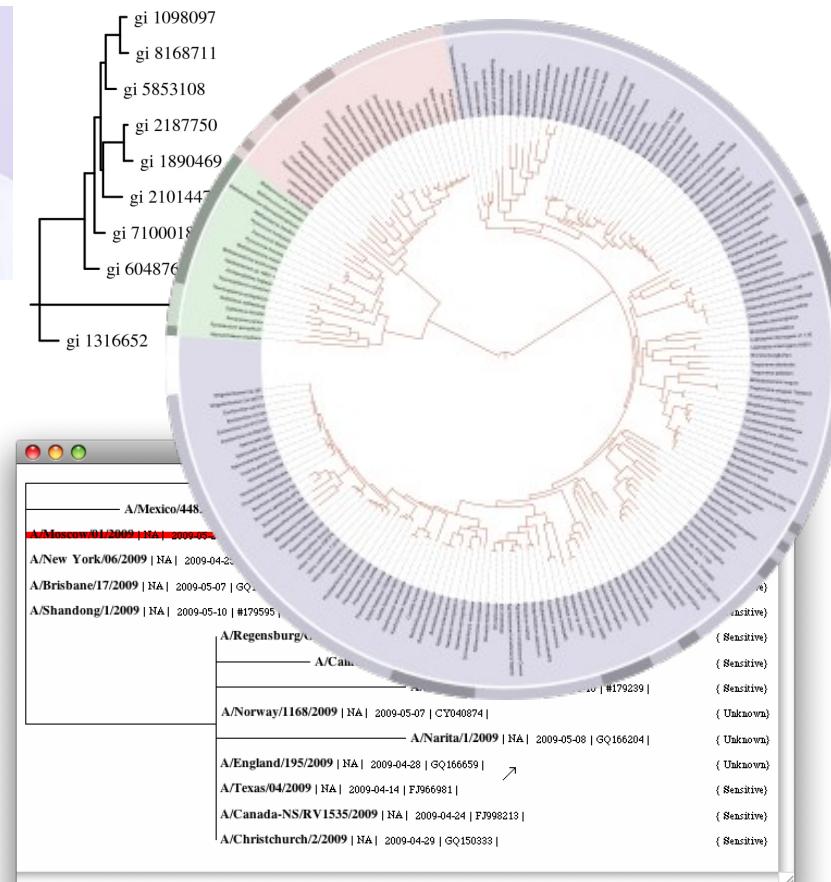
# Human efforts



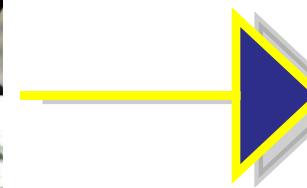
- Where does the virus come from?
- How does the virus spread?
- How does the virus evolve?

## Molecular epidemiology

- Classification of virus strains
- Tracing of transmission of a strain
- Analyses of outbreaks
- Analyses of pathogenesis of virus infection in humans



# Influenza virus data sources

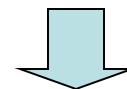


- ❖ International organizations
  - FAO
  - WHO
  - OIE
- ❖ National organizations / institutes
  - BioHealthBase
  - NCBI
  - LosAlamos

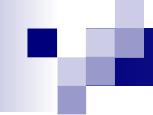
# Problems and solution

- ❖ The challenge to achieve early response is to collect critical data and to quickly process them
- ❖ Data are widely distributed geographically

WE NEED A COMPLETE AND UPDATE DATA SOURCE !!!



A GLOBAL SURVEILLANCE NETWORK ON INFLUENZA A  
USING GRID TECHNOLOGIES

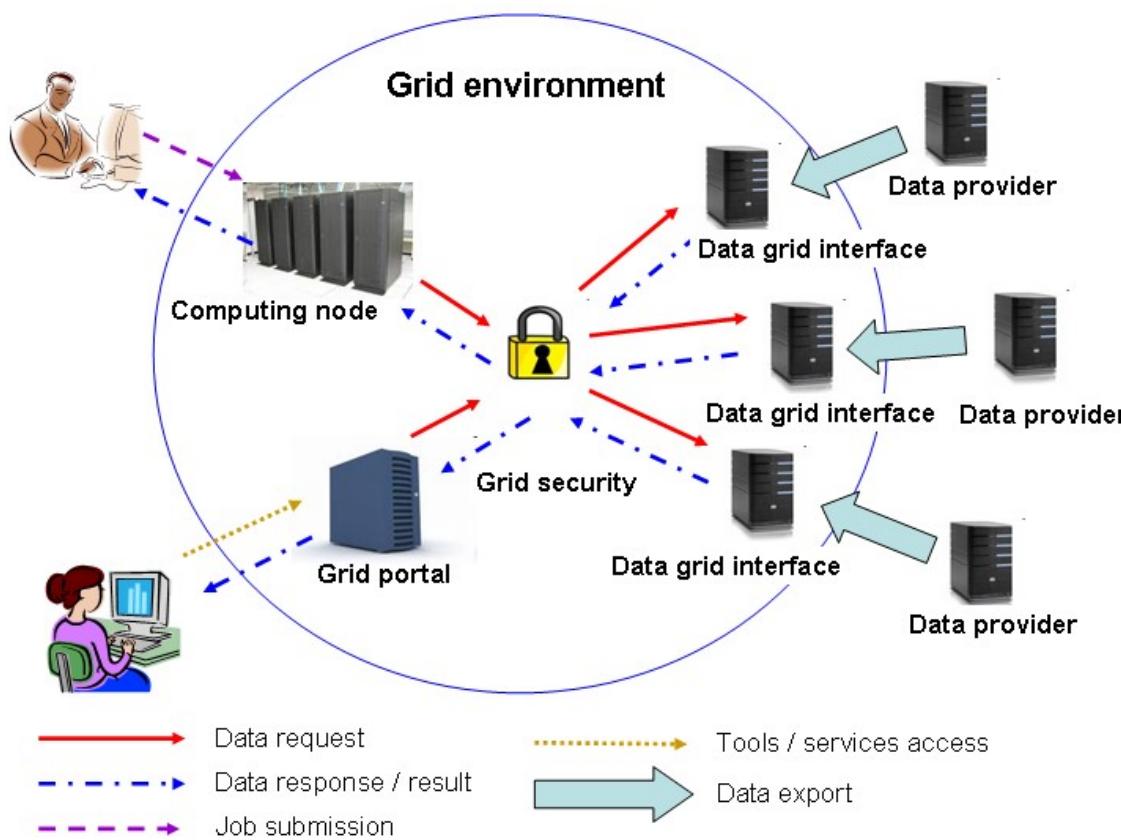


# Global Surveillance Network

# Why grid?

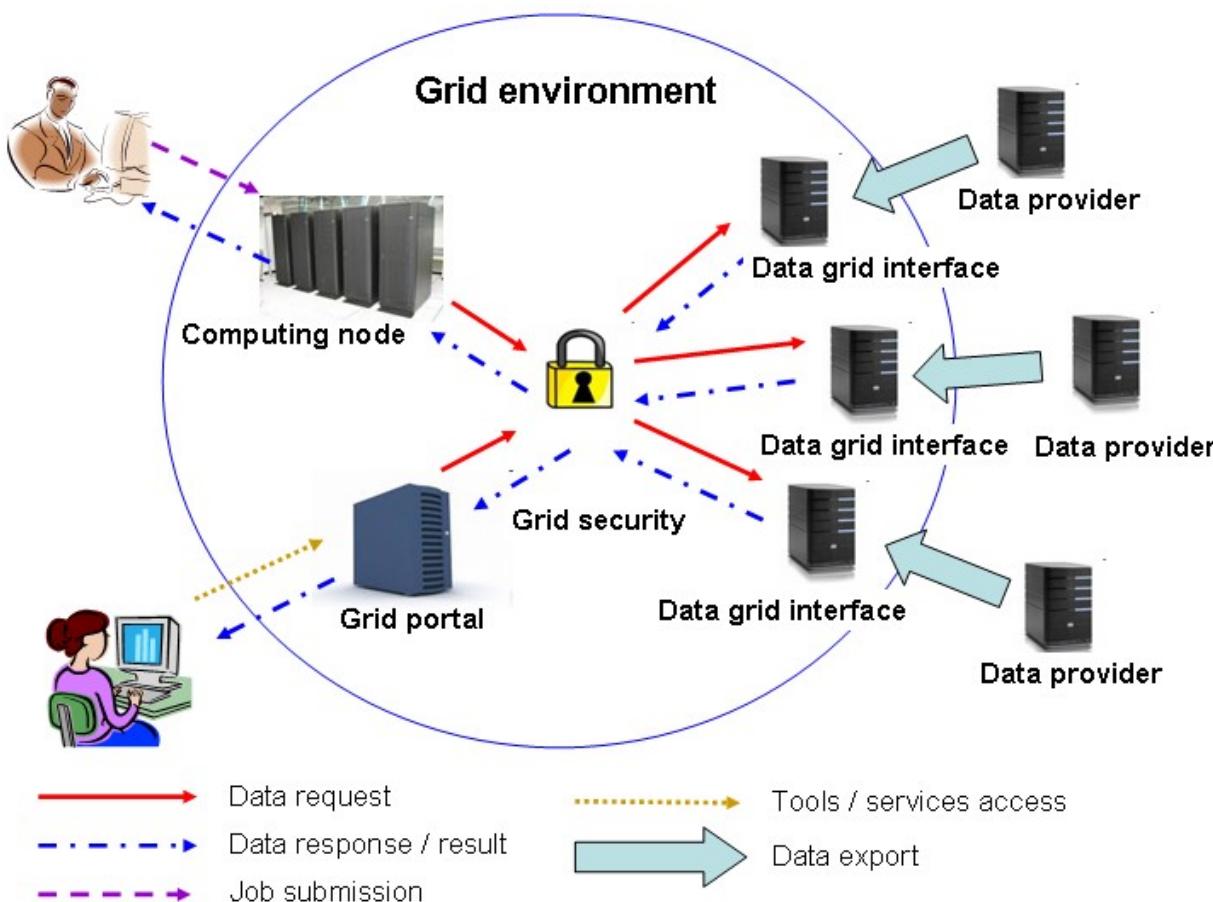
- Sharing
- Availability
- Security
- Performance

# Architecture of the surveillance network



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

# Architecture of the surveillance network



- Common molecular services will be provided on the grid
  - ❖ Data searching and retrieving
  - ❖ Phylogenetic trees
  - ❖ Estimates of the date of origin of the outbreak
  - ❖ Genetic structure of the outbreak
  - ❖ Estimates of evolutionary rate
  - ❖ Estimates of the basic reproductive number



# **NCBI database and synchornization from NCBI to AMGA**

# NCBI data

- genomeset.dat - Table with supplementary genomeset data
- influenza\_na.dat - Table with supplementary nucleotide data
- influenza\_aa.dat - Table with supplementary protein data
- influenza.dat - Table with nucleotide, protein and coding regions IDs
- influenza.fna - FASTA nucleotide
- influenza.cds - FASTA coding regions
- influenza.faa - FASTA protein

<ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/>

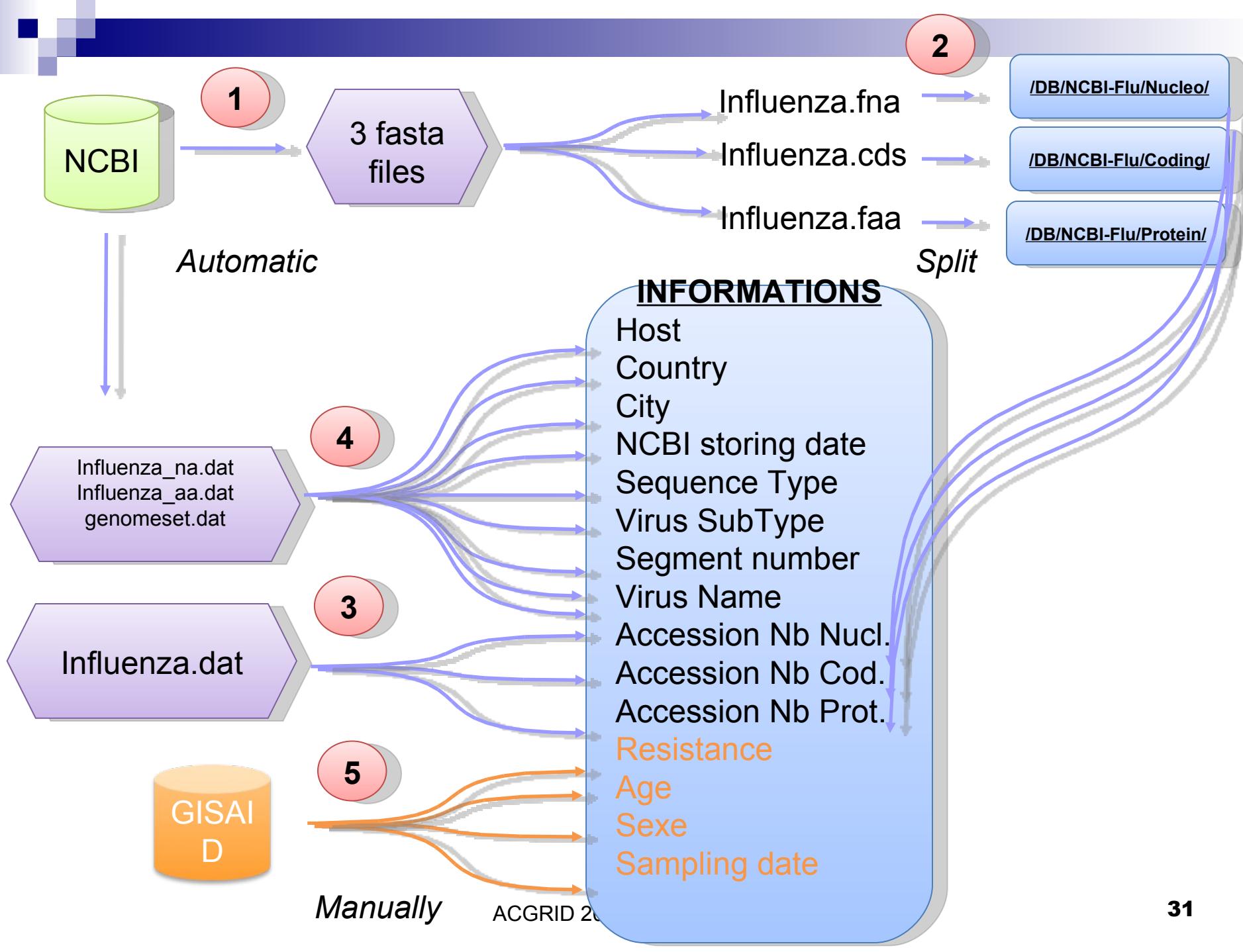
Update daily at 9 am (GMP)

# From NCBI to AMGA

- Steps between NCBI and AMGA are :

- Download the 3 fasta files (nucleotide, coding region, protein)
  - Split each files (All sequences will be separated from each other)
  - Collect IDs in the table of correspondence (influenza.dat)
  - Collect extra informations in the 3 informative files
  - Create the corresponding files (Ex: /databases/NCBI-flu/nom\_du\_fichier)
  - Create the corresponding AMGA entries

- Each fasta file will correspond to a directory in AMGA
- Each sequence will correspond to a file in a specific directory in AMGA



# Synchronization

```
>ABV25634
MKAILLVLLCAFAATNADTLCIGYHANNSTDVTVDKNTVTHSVNLLEDSHNGKLCRGGIAPLQLG
KCNIAGWLLGNPEC DLLTVSSWSYIVETSNSDNGTCYPGFIDYEELREQLSSVSSFEKFEIFPKTSSW
PNHETTRGVTAACP YAGASSFYRNLLWLVKKENSYPKLSKS YVNNKGKEVLV LWGVHHPTSTDQOSLYQ
NADAYVSVGSSKYD RRFTEIAARP KVRGQAGR MNYYWTLL EPGDTITFEATGNLVAPRYAFALNRGSES
GIITS DAPVHD CDTK CQT PHGAIN SLPF QNIHP VTIGECP KVKST KLRM VGLRN I P S I QSR G LFG A I
AGFIEGGWTGLIDGWGYHHQNGQGSGYAADQKSTQNAIDGITNKVN SIEKMNTQFTVVGKFNNLERR
IKNLNKKVDDGF LDVWTYNAELLV LLENERTLDFHD SNVKN LYEKAR SQRN NAK EIGNGC FEFYHKCDD
ACMESVRNGTYDYPKYSEESKLNREEIDGVKLESMMVYQILAIYSTVASSLVLLVSLGAISFWMC SNGSL
QCRICI
```

## FTP NCBI

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

 Vers un rép. de plus haut niveau

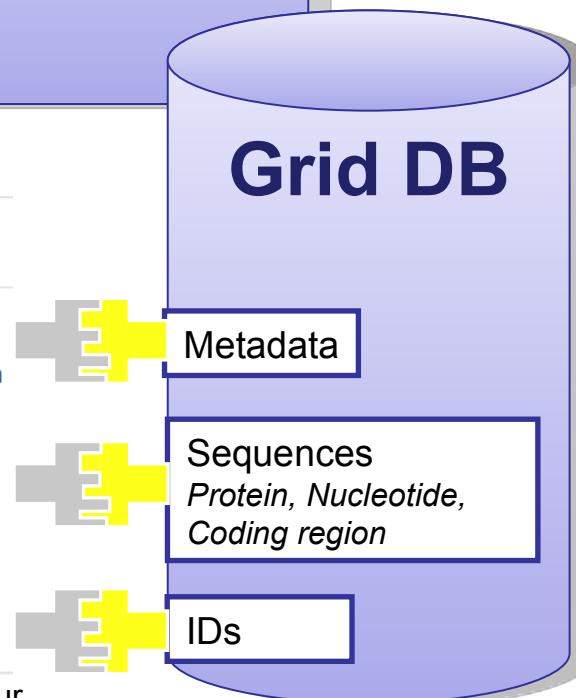
### Nom

-  genomeset.dat
-  influenza.cds
-  influenza.dat
-  influenza.faa
-  influenza.fna
-  influenza\_aa.dat
-  influenza\_na.dat



### Taille    Dernière modification

105 KB	16/07/09	08:04:00
5 KB	16/07/09	08:04:00
42 KB	16/07/09	08:04:00
97 KB	16/07/09	08:04:00
13 KB	16/07/09	08:04:00
9 KB	16/07/09	08:04:00

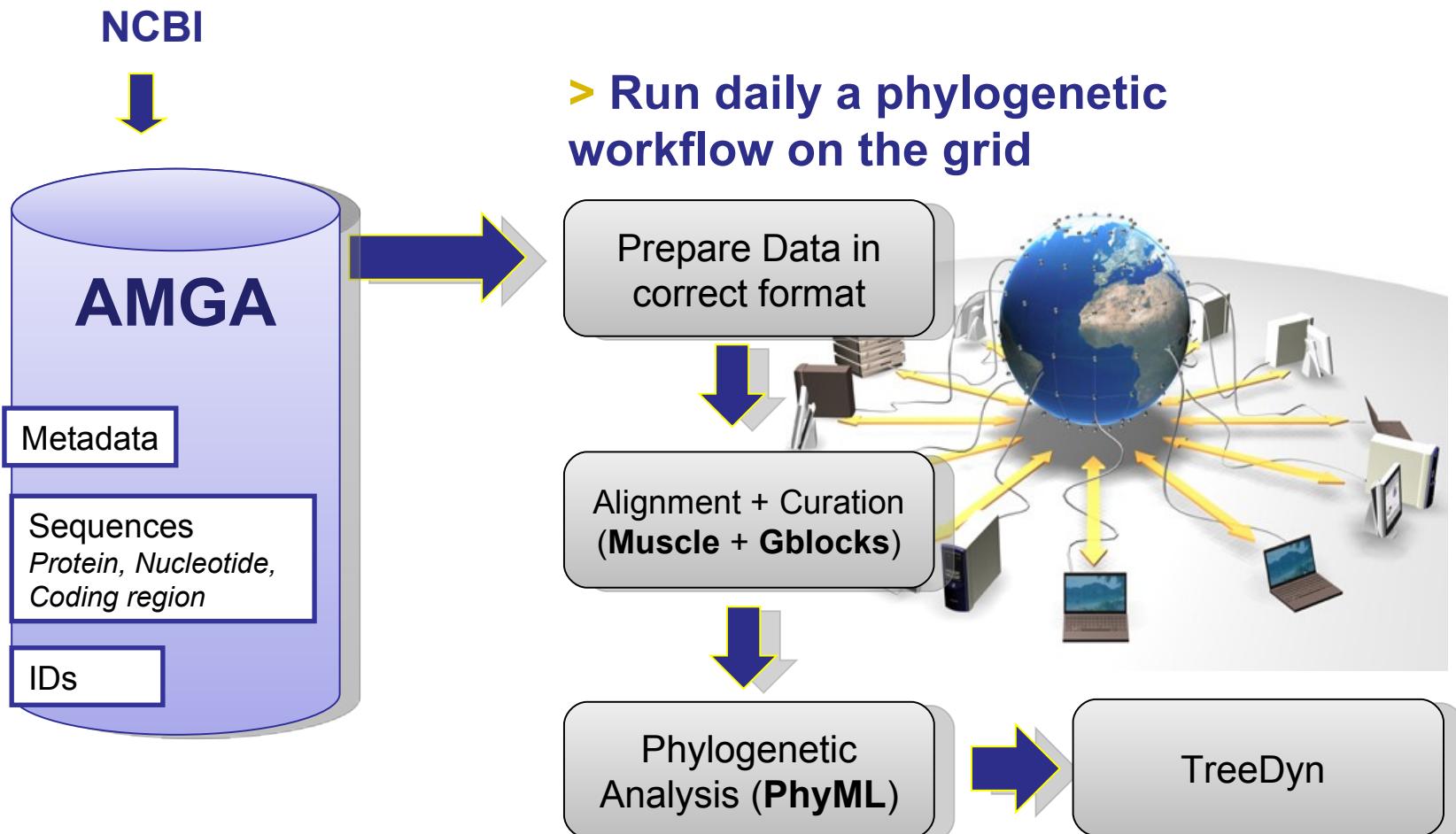


# Collections

- /Phy\_Workflow
  - /Phy\_Workflow/Task\_Monitoring
  - /Phy\_Workflow/NCBI\_Sequences
    - /Phy\_Workflow/NCBI\_Sequences/updates
    - /Phy\_Workflow/NCBI\_Sequences/coding
    - /Phy\_Workflow/NCBI\_Sequences/nucleotide
    - /Phy\_Workflow/NCBI\_Sequences/protein

# Phylogenetic workflow

# Phylogenetic workflow



# g-INFO website

The screenshot shows the homepage of the g-INFO website. At the top left is the logo, which includes a stylized globe icon and the text "grid-based International Network for Flu Observation". Below the logo is a navigation menu with the following items: Home, Flu Virus, Databases, Analysis, Results, and Contact. The "Home" item is highlighted with a teal background. The main content area features a large, abstract graphic of a network or molecular structure in blue and green. To the right of the graphic, the word "g-INFO: Home" is displayed in a large, bold, dark brown font. Below this title, there is a paragraph of text about the emergence of diseases like SARS and H1N1, followed by a detailed explanation of the challenges posed by emerging infectious diseases and the role of international collaboration and grid technologies. At the bottom, there is another paragraph about integrating existing data sources into a global surveillance network.

**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.

# g-INFO website

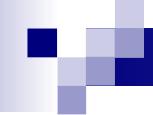
The screenshot shows the g-INFO website interface. On the left, a sidebar menu includes Home, Flu Virus, Databases, Analysis, Results (which is selected), and Contact. Below the menu is a dropdown menu for 'Arbre' (Tree) containing options: tree diag, cladogram, phenogram, eurogram, curvogram, and swoopogram. To the right of the menu is a date selector showing a calendar for September 2009. A yellow arrow points from the text 'Select branching diagram' to the 'tree diag' option in the dropdown. Another yellow arrow points from the text 'Select your date' to the calendar. The main content area is titled 'g-INFO: Results'. It displays a table of influenza sequences with columns for ID, Host, Location, Date, and various access numbers (FJ, GQ, EP). The last two columns show 'Sensitive' and 'Resistant' status. A yellow arrow points from the text 'Annotated tree with last sequences colored' to the table. The table contains the following data:

ID	Host	Location	Date	FJ	GQ	EP	Sensitive	Resistant
FJ966084	A/California	04/2009	2009-04-01	FJ966084		EPI176472	Sensitive	
FJ966081	A/Texas	04/2009	2009-04-14	FJ966081		EPI176505	Sensitive	
FJ984953	A/Regensburg	Germany	01/2009		09166204	A/Norita/1/2009   2009-05-08 GQ166204   EPI179239   Unknown	Unknown	
CY040874	A/Norway	1168	2009		09166204	2009-05-07 CY040874   EPI179321   Unknown	Unknown	
QQ166559	A/England	195	2009		09159333	A/England/195/2009   2009-04-29 GQ159333   EPI178516   Unknown	Unknown	
FJ986213	A/Christchurch	2/2009			09159333	2009-04-29 FJ986213   EPI178286   Sensitive	Sensitive	
QQ365445	A/Osaka	180	2009		09365445	A/Osaka/180/2009   2009 GQ365445   EPI184961   Resistant	Resistant	
FJ966082	A/Denmark	529	2009	FJ966082		EPI190216   Resistant	Resistant	
FJ984950	A/Mexico	4482	2009		09149670	A/Mexico/4482/2009   2009-04-14 GQ149670   EPI176588   Sensitive	Sensitive	
FJ984340	A/New	York	06/2009		09184629	A/New York/06/2009   2009-04-21 FJ984340   EPI177375   Unknown	Unknown	
QQ160610	A/Brisbane	17/2009			09160610	A/Brisbane/17/2009   2009-05-07 GQ160610   EPI179072   Sensitive	Sensitive	
QQ200288	A/Shandong	1/2009			09200288	A/Shandong/1/2009   2009-05-10 GQ200288   EPI179955   Sensitive	Sensitive	
QQ132156	A/Canada-AB	RV1532	2/2009		09132156	A/Canada-AB/ RV1532/2009   2009 GQ132156   EPI185352   Sensitive	Sensitive	
FJ980215	A/Canada-ON	RV1527	2/2009		09132156	A/Canada-ON/RV1527/2009   2009-04-24 FJ980215   EPI183120   Sensitive	Sensitive	
QQ220730	A/Hyogo	1/2009			09220730	A/Hyogo/1/2009   2009-05-17 GQ220730   EPI186248   Unknown	Unknown	
QQ220736	A/Amagasaki	C-1/2009			09220736	A/Amagasaki/C-1/2009   2009 GQ220736   EPI180735   Sensitive	Sensitive	
QQ220734	A/Osaka	C-1/2009			09220734	A/Osaka/C-1/2009   2009-05-16 GQ220734   EPI180733   Sensitive	Sensitive	

# Demo of phylogenetic workflow

# Conclusions

- ❖ Vision of a global surveillance network for influenza A
  - Complete and update data source
  - Molecular services
- ❖ This surveillance network is different from others by using grid technology
- ❖ It will need a very active collaboration



# Questions?

# Hands-on

# AMGA queries

- Start `mdclient`, `cd` to `/tung/ncbi` and try the following queries on the table `metadata`:
  - Count number of sequences of virus H1N1 in the database
  - Find sequences of virus H1N1 that host = human
    - Verify if the last returned sequence has host = human by using `getattr`
  - Find sequences of virus H1N1 that host = human and country = malaysia
    - Verify the last returned sequence
  - Find sequences of virus H1N1 that host = human, country = USA and city = California (*city is often provided in virus name, use function `like` to get it*)
    - Verify the last returned sequence

# Create AMGA tables

- Start **mdclient**, cd to /tung/users, create a directory, name it after your username, for example: **/tung/users/test**
- then cd to your directory, for example: cd /tung/users/test
- Create the following tables:
  - **metadata** with attributes: host: varchar(30), country: varchar(30), year: varchar(20), subtype: varchar(10), segnum: int, seqlen: int, virusname: varchar(200), an\_proteins: varchar(100), an\_codings: varchar(100), age: varchar(50), sex: varchar(50)
  - **nucleotide** with attributes: sequence: text
  - **protein** with attributes: sequence: text
  - **coding** with attributes: sequence: text
  - **updates**: no attributes
- Commands to use: **createdir**, **addattr**

# Run update script (1)

- Copy update scripts from `/home/kualalumpur30/ncbi.tar.gz` to your home directory
- Extract scripts: `tar -xzf ncbi.tar.gz`
  - `/home/~yourname/ncbi`
    - `fasta`
    - `downloadncbi.sh`
    - `data_handle.sh`
    - `extract_nuc.sh`
    - `extract_pro.sh`
    - `extract_code.sh`
    - `run_nuc.sh`
    - `run_pro.sh`
    - `run_cod.sh`
    - `run_update.sh`
- `cd` to `ncbi` directory

# Run update script (2)

- Download updates from NCBI

- `./downloadncbi.sh  
ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/updates/2009-11-08`

- Run updates:

- `./run_nuc.sh 081109`
  - `./run_pro.sh 081109`
  - `./run_cod.sh 081109`

- Verify these update commands

- Start mdclient, cd to your directory then list the directories of the date you specified
    - `ls updates/081109/nucleotide`
    - `ls updates/081109/protein`
    - `ls updates/081109/coding`

- Try another date

# Run simple workflow (1)

- Copy workflow script from `/home/kualalumpur30/workflow.tar.gz`, extract it to your directory
  - `/home/~/yourdir/workflow`
    - `workflow.sh`: Script that runs a simple workflow
    - `muscle`: Program to align sequences
    - `phyml`: Program to construct phylogenetic trees
- cd to workflow then run `workflow.sh` with 4 parameters:
  - `./workflow.sh Human H1N1 "Viet Nam" 2005`
- You should get the following files as results:
  - `data.aligned.fasta`
  - `data.aligned.fasta.phy`
  - `data.aligned.fasta.phy_phyml_lk.txt`
  - `data.aligned.fasta.phy_phyml_boot_stats.txt`
  - `data.aligned.fasta.phy_phyml_stat.txt`
  - `data.aligned.fasta.phy_phyml_boot_trees.txt`
  - `data.aligned.fasta.phy_phyml_tree.txt`

# Run simple workflow (2)

- Go to <http://www.phylogeny.fr/>
  - Choose “Online Programs” > TreeDyn
  - Paste content of data.aligned.fasta.phy\_phyml\_tree.txt to textbox  
then submit

