# Dude, where's my sample?
## The inner workings of the PANGEA database

Tanya Golubchik


# OVERVIEW

Data flow within PANGEA

For the impatient: PANGEAdb in one slide

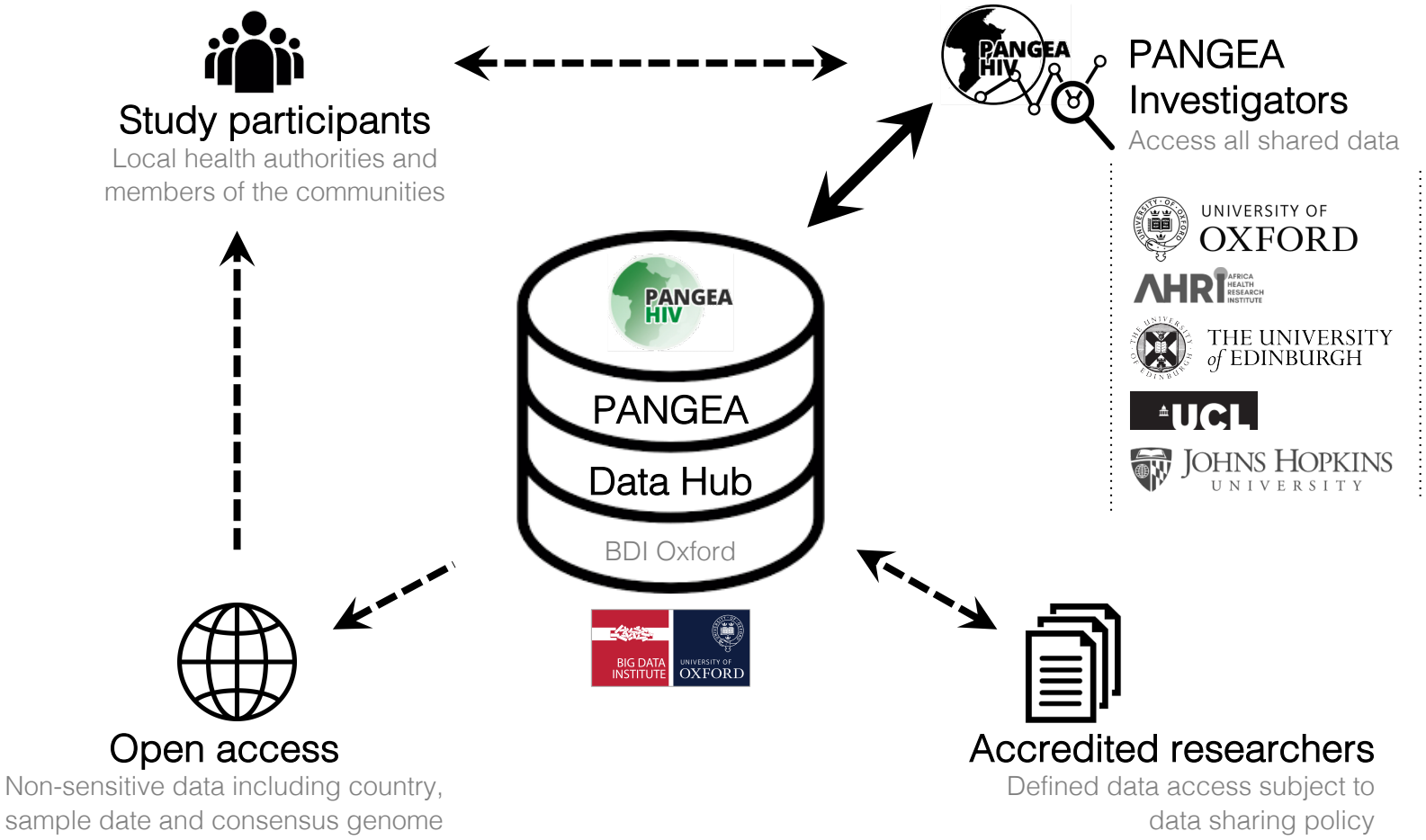
In more detail: How PANGEAdb is organised*

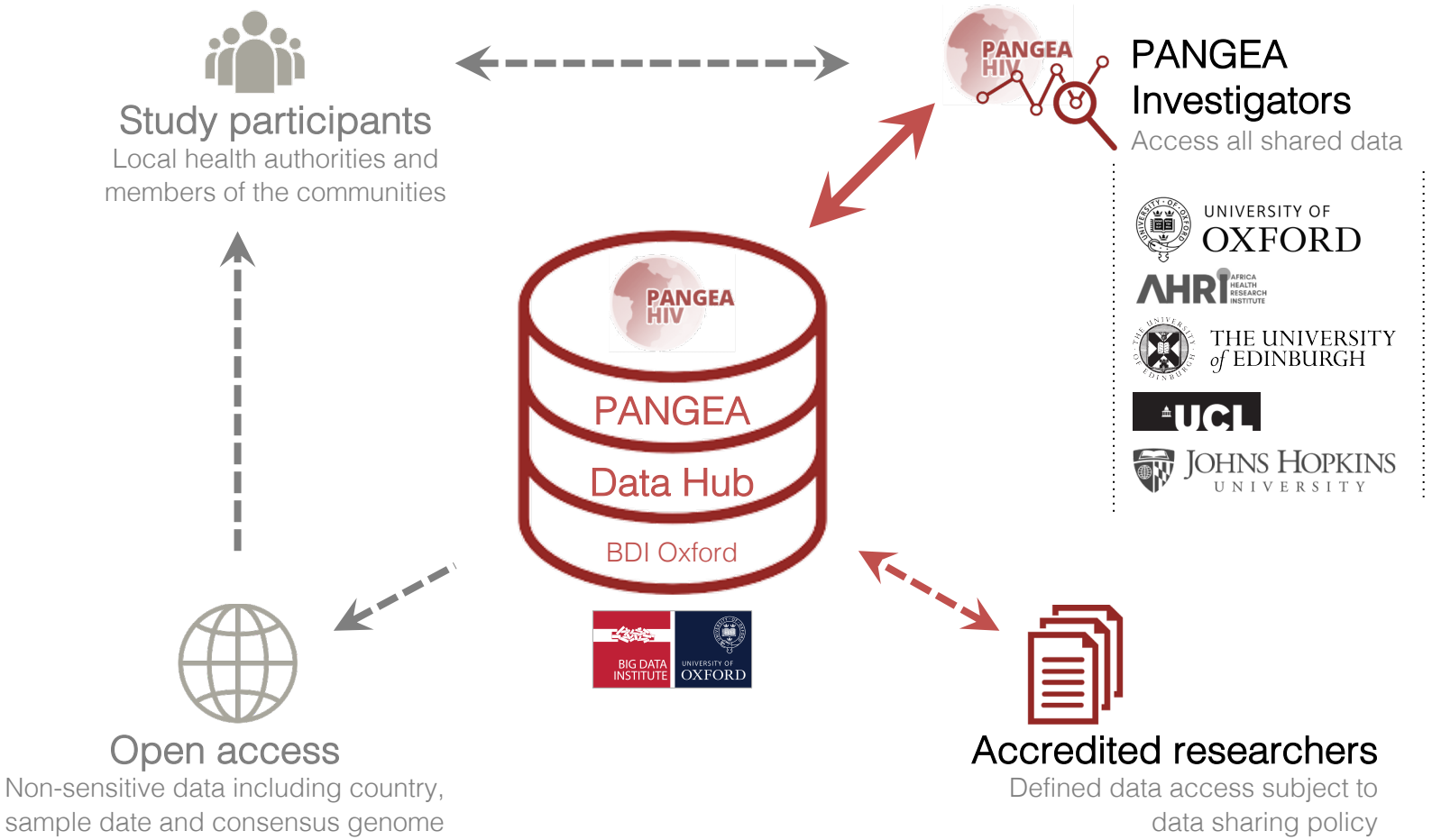
* *AKA there is no "sample"*

What happens when you request data from PANGEA?

What happens when you send samples for sequencing?

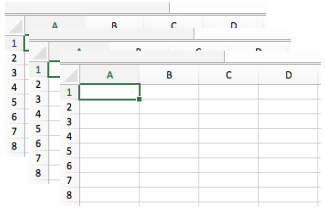
Summary and questions


Study participants
Local health authorities and members of the communities
PANGEA HIV
PANGEA Investigators
Access all shared data
UNIVERSITY OF OXFORD
AHRI AFRICA HEALTH RESEARCH INSTITUTE
THE UNIVERSITY of EDINBURGH
UCL
JOHNS HOPKINS UNIVERSITY
PANGEA HIV
PANGEA
Data Hub
BDI Oxford
BIG DATA INSTITUTE
UNIVERSITY OF OXFORD
Open access
Non-sensitive data including country, sample date and consensus genome
Accredited researchers
Defined data access subject to data sharing policy

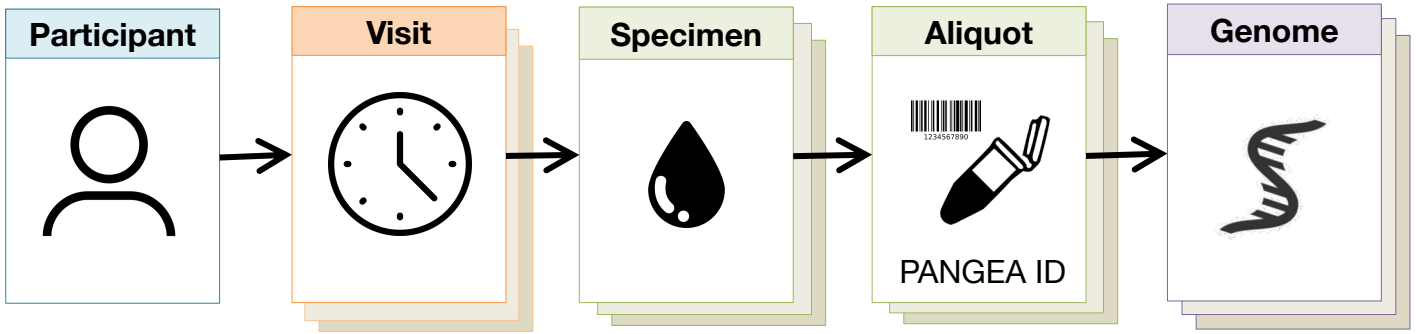

Study participants
Local health authorities and members of the communities
PANGEA HIV
PANGEA Investigators
Access all shared data
UNIVERSITY OF OXFORD
AHRI AFRICA HEALTH RESEARCH INSTITUTE
THE UNIVERSITY of EDINBURGH
UCL
JOHNS HOPKINS UNIVERSITY
PANGEA HIV
PANGEA
Data Hub
BDI Oxford
BIG DATA INSTITUTE
UNIVERSITY OF OXFORD
Open access
Non-sensitive data including country, sample date and consensus genome
Accredited researchers
Defined data access subject to data sharing policy

# PANGEAdb IN ONE SLIDE

- Relational database (MySQL)
  - **Rows** of related information are grouped into **tables** which together make up the **schema**
  - Each row is labelled by a unique identifier (primary **key**)
  - These keys link tables together



- In PANGEAdb the schema **follows real-life flow of information:**




**Who?**
Time-invariant individual data, eg. Year of Birth

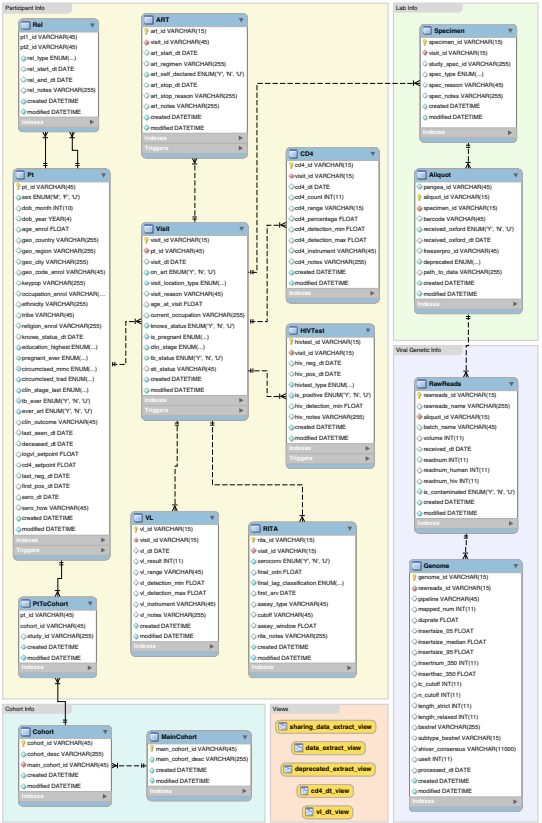
**When?**
Time-specific data, eg:
Survey round

**What?**
Biological material
Eg. plasma, proviral DNA

**Which tube?**
Barcoded tube as received
PANGEA ID

**What is the best sequence?**
Raw data
Assembled genome
Minor variants
Quality statistics
….

# IN MORE DETAIL



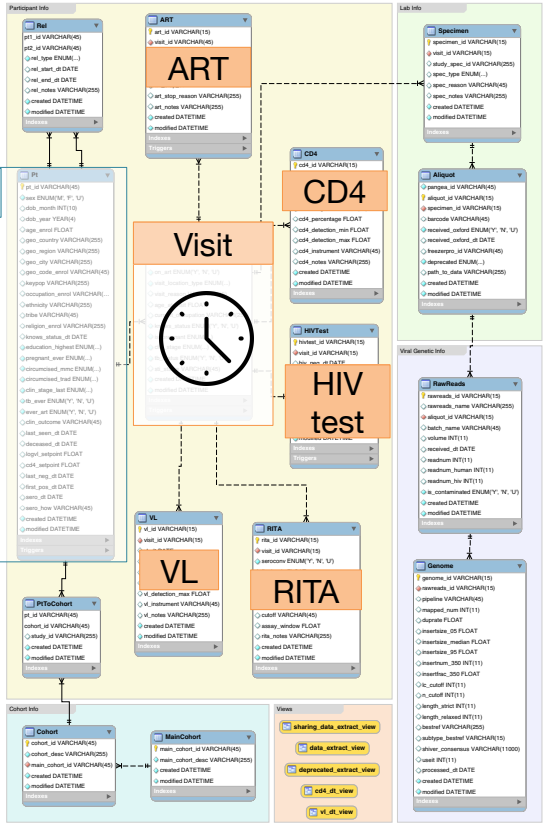
# … TOO MUCH DETAIL!

# OK, A *LITTLE* MORE DETAIL

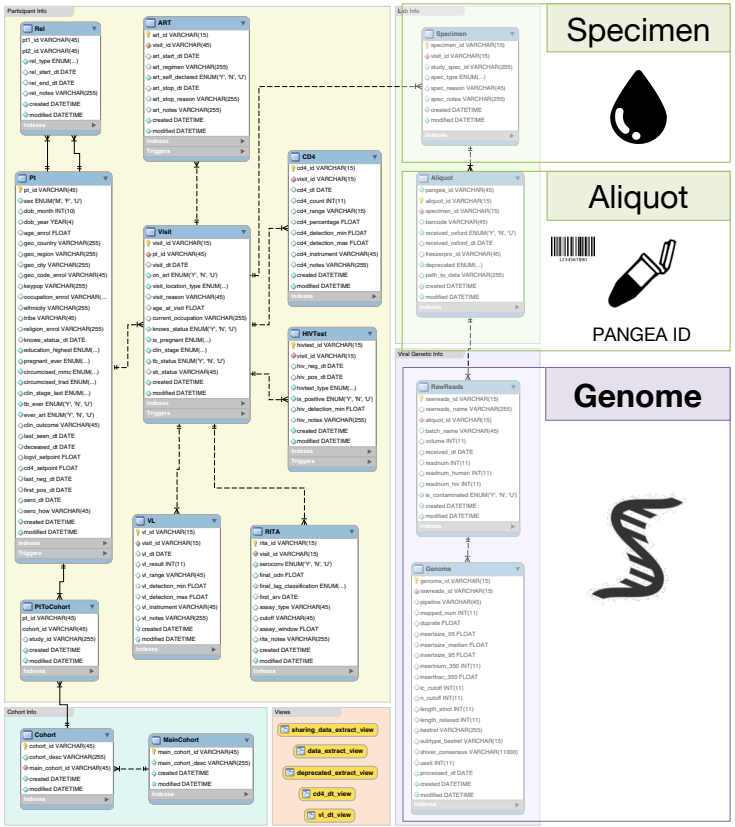
Participant

pt_id
gender
dob_year
dob_month
geo_country
ethnicity
keypop
art_ever
…etc

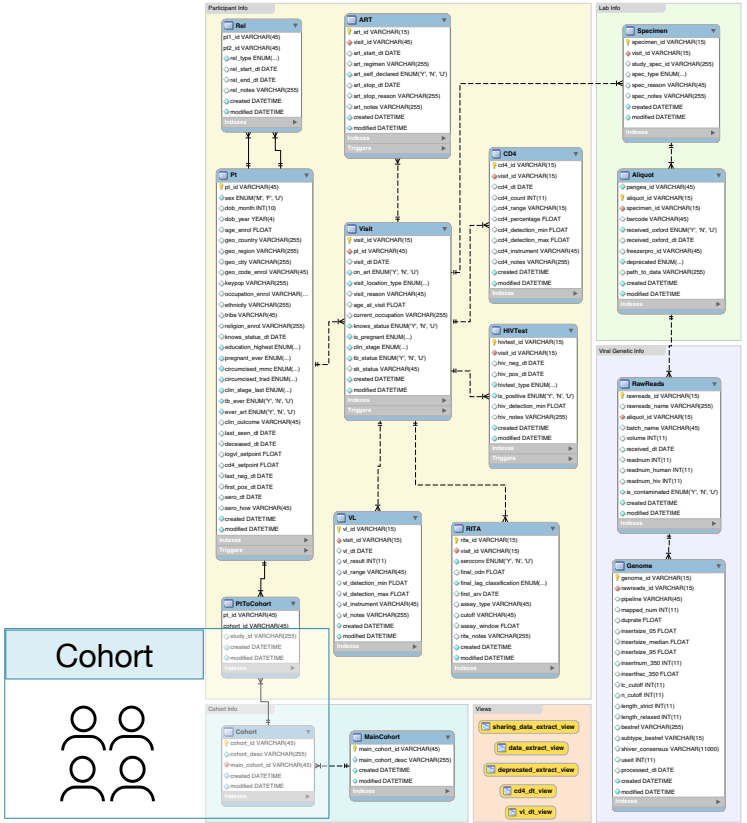
Calculated fields:
first_pos_dt
last_neg_dt
…



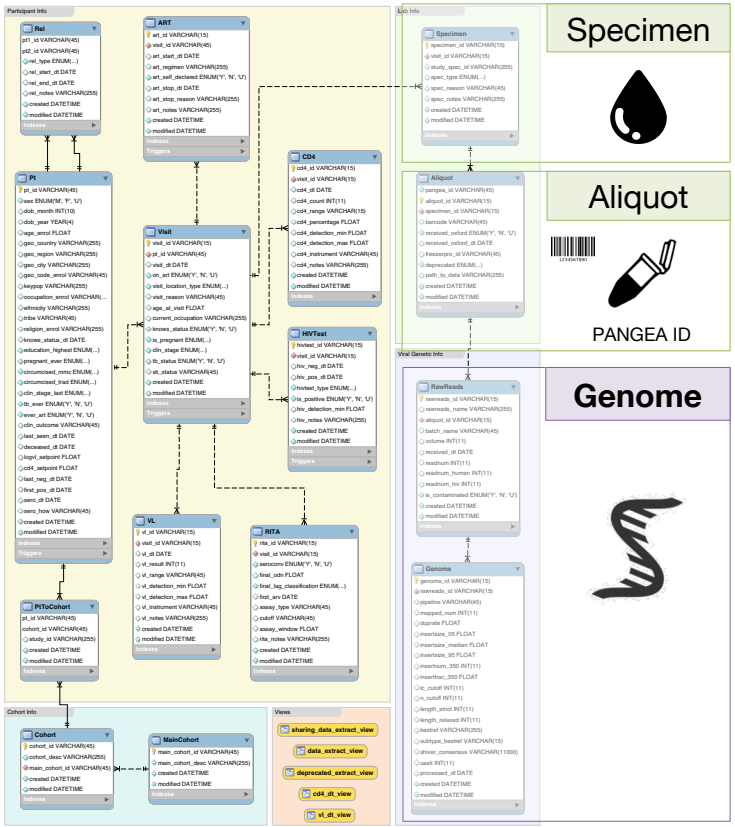
# LAB DUPLICATES ARE MODELLED



# MULTIPLE COHORTS ARE ALSO MODELLED



# WHAT IS A **SAMPLE**?

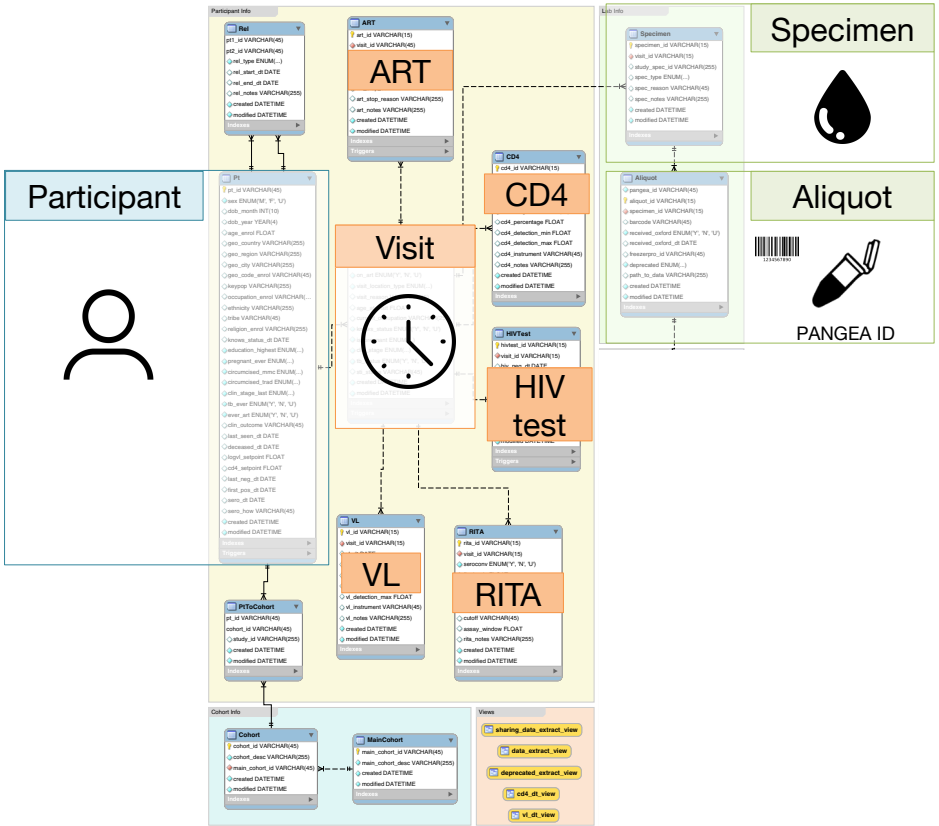



Is it a blood draw?

Is it what's in the tube?

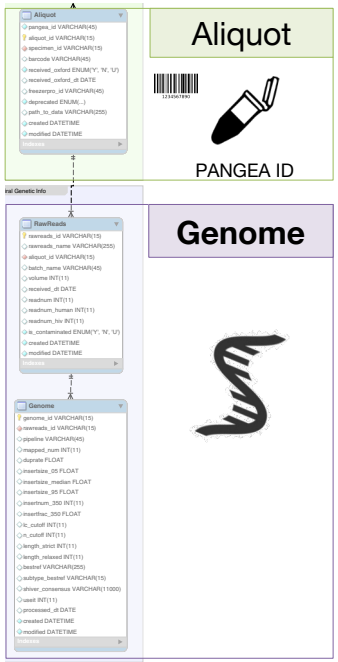
Is it a sequence?

# WHAT YOU KNOW

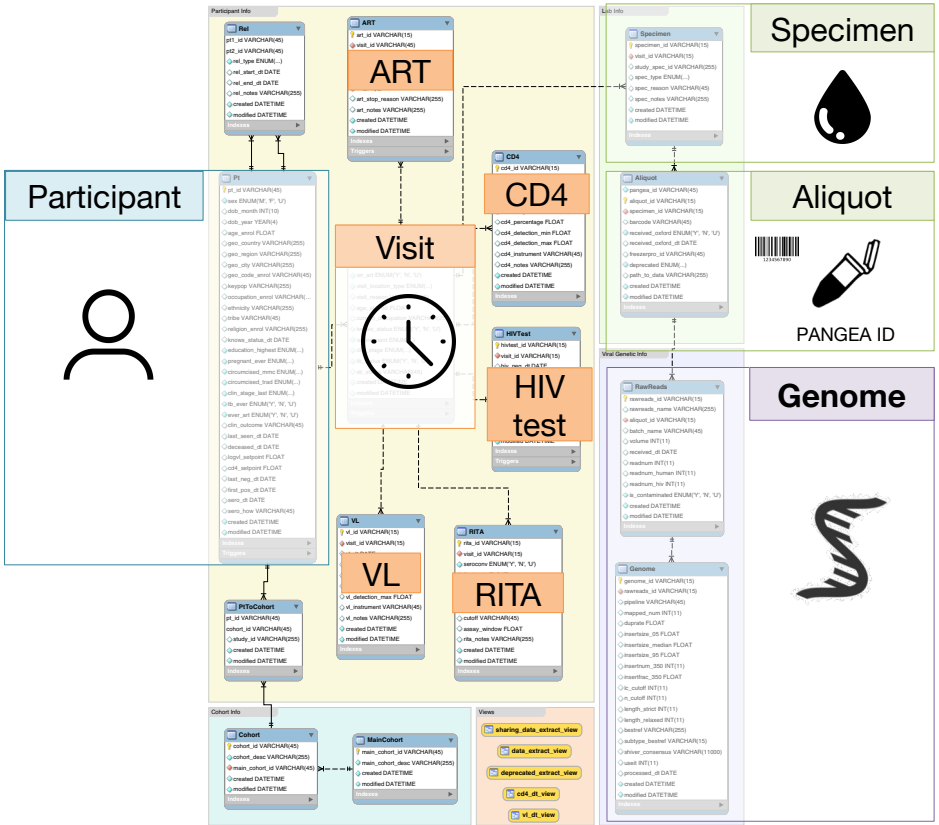


# WHAT THE LAB KNOWS

# WHAT THE BIOINFORMATICIAN KNOWS



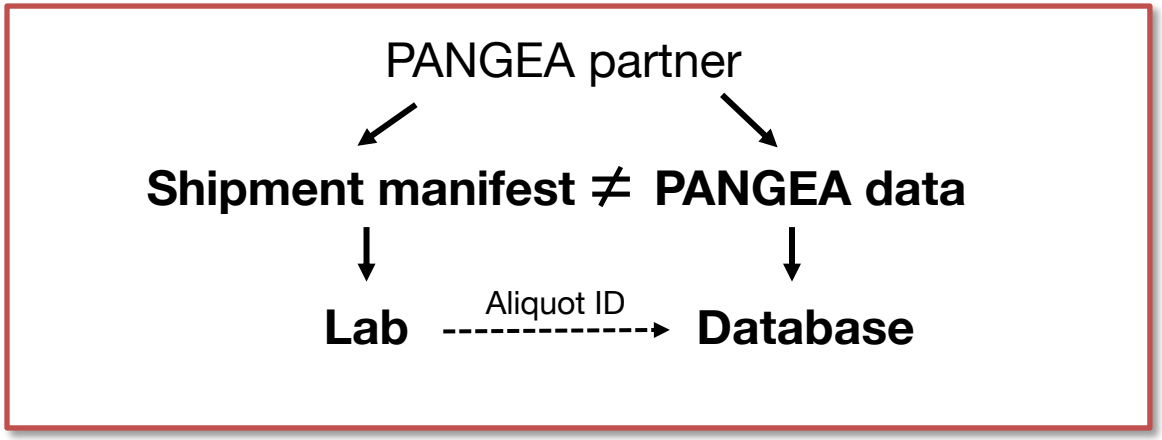
# PANGEADB: THE WHOLE PICTURE



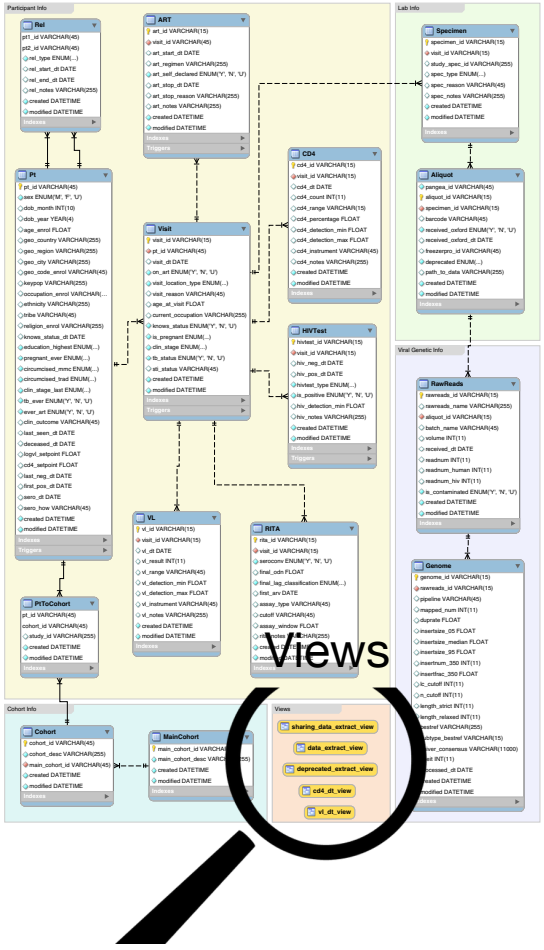
# WHAT HAPPENS WHEN YOU SEND US SAMPLES?

- The PANGEAdb team will send you a **data request** specifying the fields we need
  - PANGEA partners can return the information as flat CSV files, or structured tables
  - We use this information to assign **PANGEA IDs** to each barcoded tube that the lab will process
  - For new projects: we will map any new fields onto the database and extend the schema if necessary

- **Aliquots** are received by the Oxford **lab**
  - Shipment manifest is checked against PANGEAdb to ensure all barcodes have matching data
  - Participant data **must exist** for us to proceed with sequencing



# What happens when you request data?

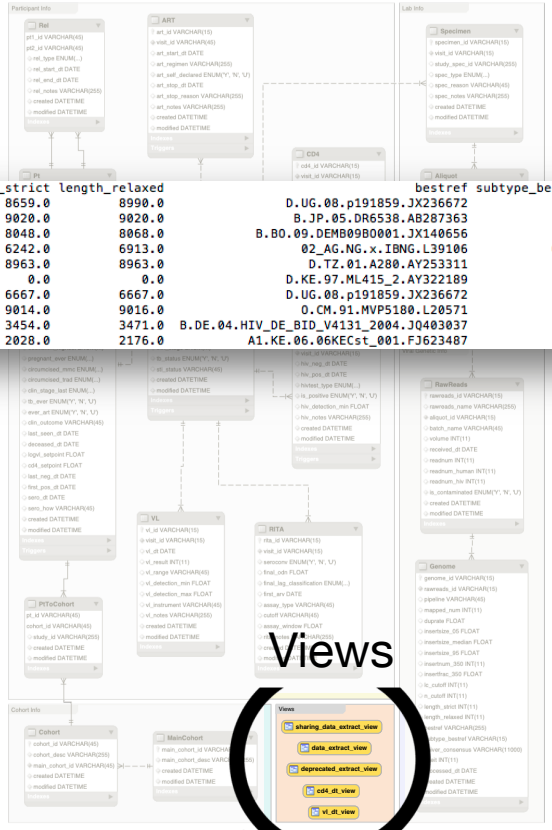



- **Views** are pre-defined queries
- The default **data sharing view** combines basic information from many tables
  - Outputs a “spreadsheet”
- PANGEA collaborators can request a full **global join** of all data we hold (except highly sensitive data)

# What happens when you request data?

Data sharing example:

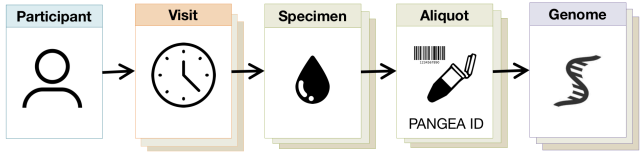
| sex | age_enrol | geo_country | ever_art | ... | length_strict | length_relaxed | bestref | subtype_bestref | shiver_consensus |
|---|---|---|---|---|---|---|---|---|---|
| F | 66.4 | Uganda | U | ... | 8659.0 | 8990.0 | D.UG.08.p191859.JX236672 | D | GCTAGCTAGGGAACCCACTGCTTAAGCCTCAATAAAGCTTGCCTTG... |
| U | NaN | Uganda | U | ... | 9020.0 | 9020.0 | B.JP.05.DR6538.AB287363 | B | GCTAACTAGGGAACCCACTGCTTAAGCCTCAATAAAGCTTGCCTTG... |
| U | Na | Uganda | U | ... | 8048.0 | 8068.0 | B.BO.09.DEMB09B0001.JX140656 | B | ???AGCTAGGGAA :TTAAGCCTCAATAAAGCTTGCCTTG... |
| F | 42. | Uganda | U | ... | 6242.0 | 6913.0 | 02_AG.NG.x.IBNG.L39106 | 02_AG | ???????AGGGAA :TTAAGCCTCAATAAAGCTTGCCTTG... |
| F | 47. | Uganda | Y | ... | 8963.0 | 8963.0 | D.TZ.01.A280.AY253311 | D | GCTAGCTAGGGAA :TTAAGCCTCAATAAAGCTTGCCTTG... |
| F | 42. | Uganda | Y | ... | 0.0 | 0.0 | D.KE.97.ML415_2.AY322189 | D | ????????????? ??????????????????????????... |
| M | 41. | Uganda | U | ... | 6667.0 | 6667.0 | D.UG.08.p191859.JX236672 | D | GCTGGCTAAGGAA :TTAAGCCTCAATAAAGCTTGCCTTG... |
| M | 27.4 | Uganda | U | ... | 9014.0 | 9016.0 | O.CM.91.MVP5180.L20571 | O | ?????????GGAACCCACTGCTTAAGCCTCAATA?a-GCTTGCCTT... |
| M | 60.0 | Uganda | U | ... | 3454.0 | 3471.0 | B.DE.04.HIV_DE_BID_V4131_2004.JQ403037 | B | GCTAGCAAGGGAACCCACTGCTTAAGCCTCAATA???GCTTGCCTT... |
| F | 40.4 | Uganda | U | ... | 2028.0 | 2176.0 | A1.KE.06.06KECst_001.FJ623487 | A1 | ??????????????????????????????????????????????... |



# SUMMARY

- PANGEAdb is a relational database
  - Schema is made up of linked tables which are made up of rows
  - Each table can be thought of as a “spreadsheet”, with rows labelled by a primary key = unique identifier

- We model data flow from participant to HIV genome
- Basic principles:
  - Keep it simple – complexity can be built up as needed
  - Keep it real – focus on physical entitites
  - Keep it unique – allows for modelling of duplicates at any level (specimen, aliquot, sequence…)




- Shipment manifest is not data sharing

# QUESTIONS?