Topic

- Data entry experience from various participants
- IHIWS XML
- Histoimmunogenetics Markup Language (HML) to IHIWS XML
- Standardized HML report
- Reach agreements reporting Novel Polymorphisms

Kazutoyo Osoegawa, Ph.D.

HLA Lab
- Provides sample
- Generates data

→ Database →

HLA Informatics group
Needs data to analyze

HLA Lab is not familiar with HLA informatics
- Enter data “painlessly”
- Tend to “give up”

HLA informatics group
- Extract data in analyzable format

Database has been developed to accept data in analyzable format

↓

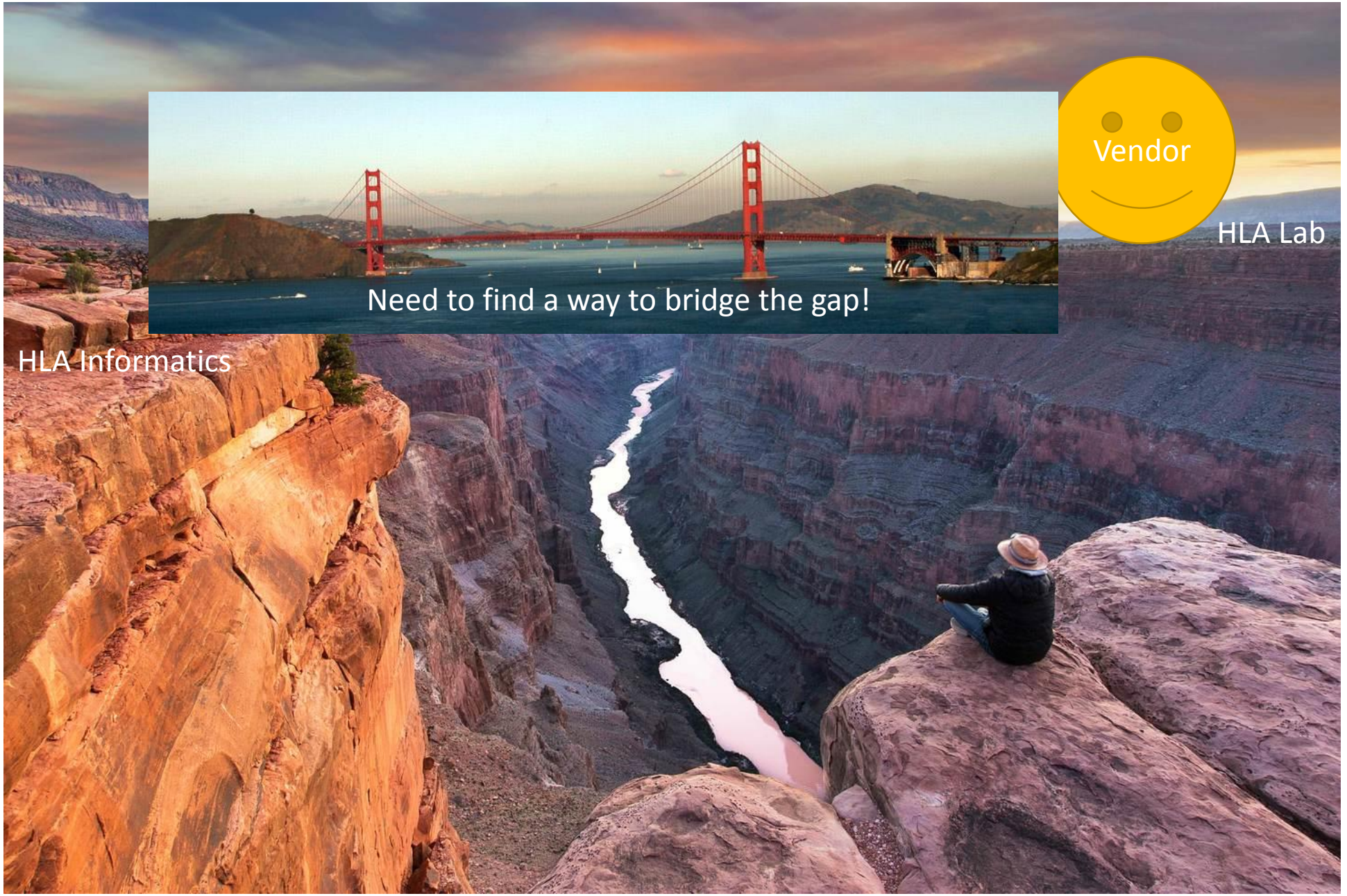
So far, “data submission extremely difficult and time consuming”

“I will likely have to re-think my participation as I did not appreciate how much work would be required to upload the data”

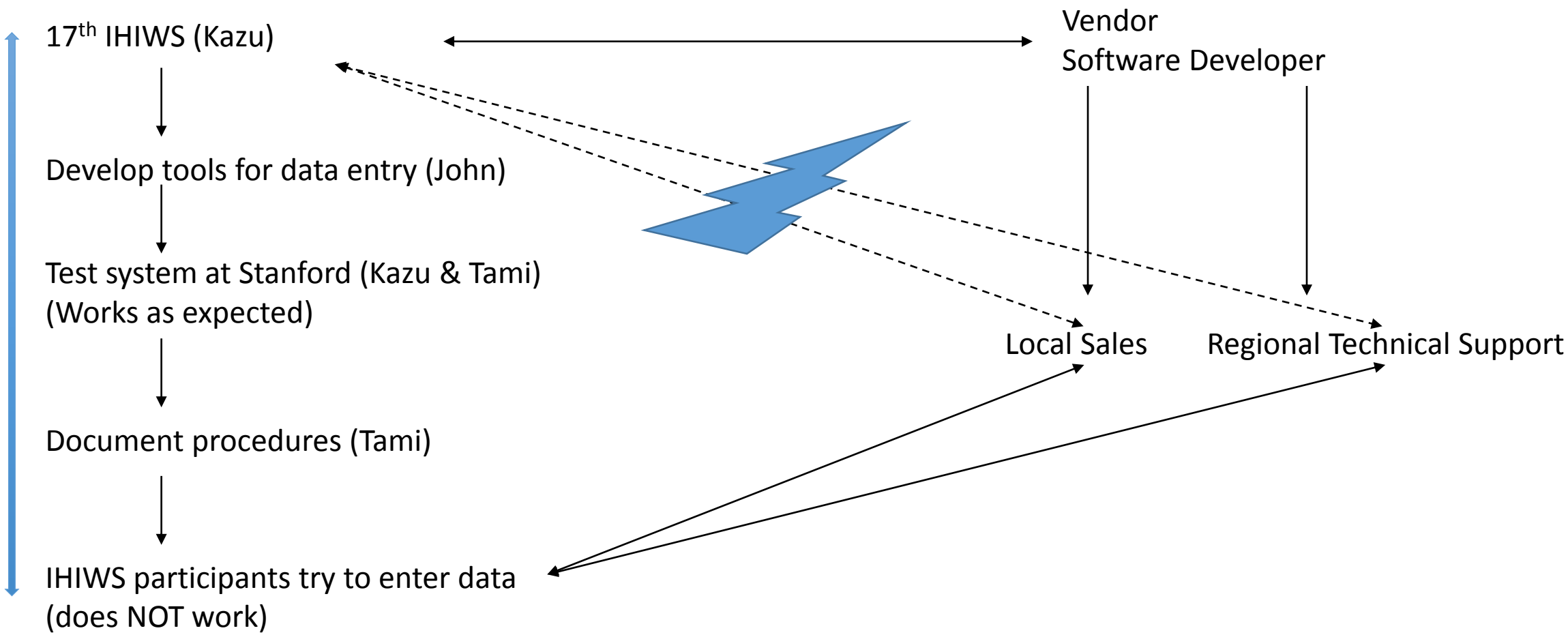
↓

We need to work with vendors to facilitate entering data into database

Why we are having this meeting?



A large gap between HLA informatics and HLA Lab?



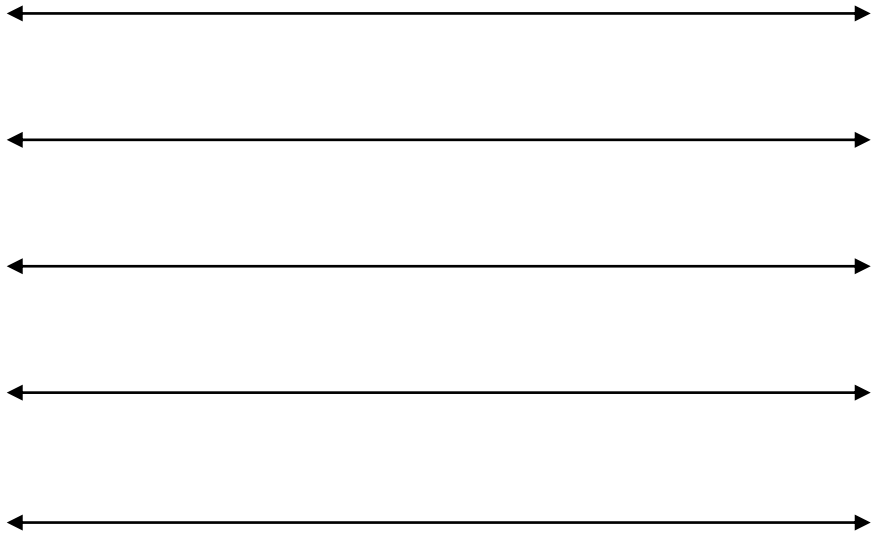
Current Communication Flow

Each vendor has their own way

17th IHIWS

IMPOSSIBLE to accept different format of output file from each vendor's system

Vendor will not likely agree on generating workshop XML format



Participant1	Vendor1
Participant2	Vendor2
Participant3	Vendor3
Participant4	Vendor4
Participant5	Vendor5

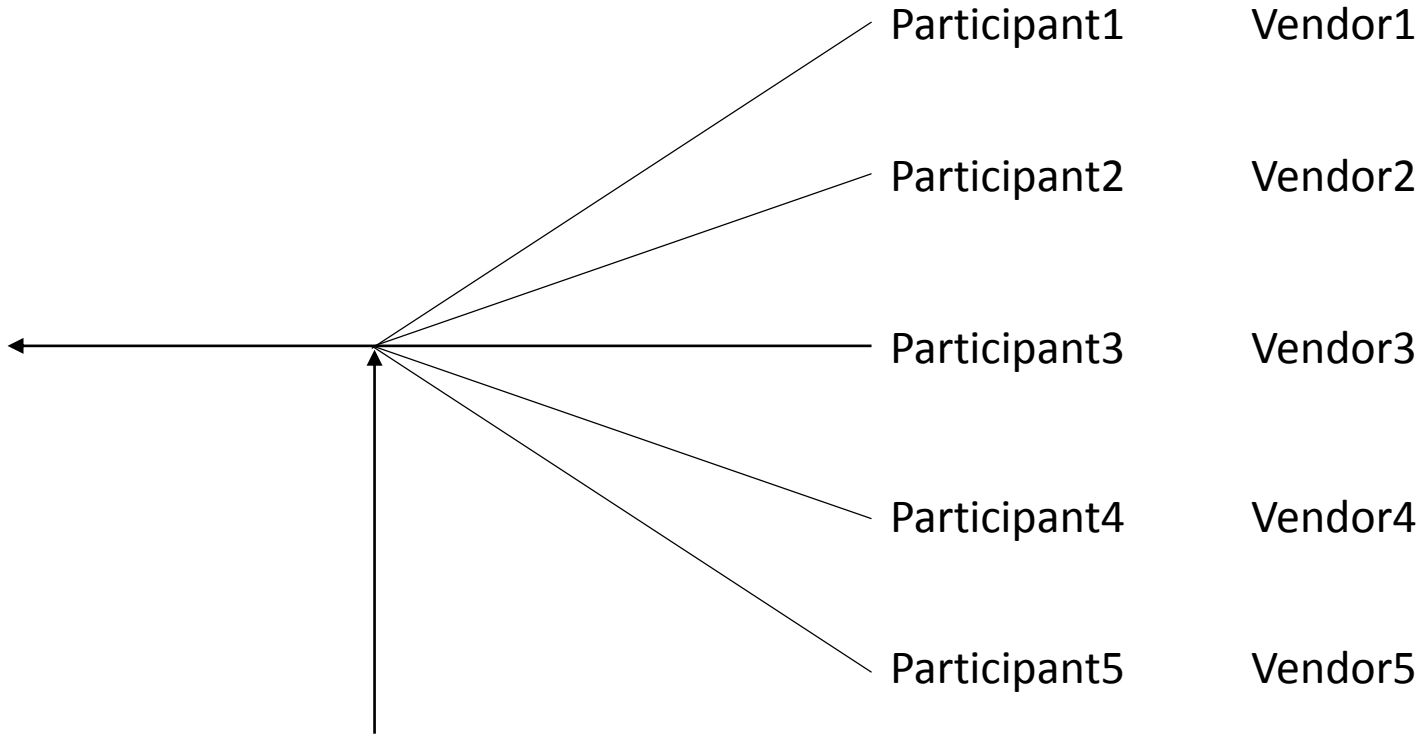
In CSV file
Output formats are different for each vendor
Output formats change at software updates!

There are many different ways to report from HLA software!

Each vendor has their own way

17th IHIWS

- Need to find a mutually beneficial and agreeable output
- Develop tools to convert HML to workshop XML



Histoimmunogenetics Markup Language (HML) or similar XML output

BUT, has HML output

If the vendor follows the HML schema, then we should be able to capture most of the data that we would like to populate

```
<IHIW_Report>
<!--[ ... ]-->
<Lab LabCode="6 character code" Lab_defined_ID="Free Text">
<Software_Applied Software_Manufacturer="Free Text" Software_Name="Free Text"
Software_Parameters="Free Text" Software_Version="Free Text" Sofware_Function="CV List"/>
<Hardware_Used Instrument_Firmware="Free Text" Instrument_Model_Number="Free Text"
Instrument_name="Free Text" Manufacturer="Free Text"/>
<Reagent_Protocol protocol_name=“Free Text” protocol_source=“Free Text”
protocol_external_identifier=“Free Text” protocol_internal_identifier=“Free Text” protocol_deviations=“Free Text” />
<Sample SampleID="Free Text">
<Genotyping Genotype_GL="GL String Formatted Genotype">
<Locus HLATyping="CV List" Alignment_Reference_DB="Free Text" BaseCalling_Reference_DB="Free Text"
Consensus_Sequence="Free Text" Feature="CV List" Locus_name="CV List" MeanReadDepth="Free Text"
DataFileLoc="URL or Filepath" PhasingGroup="Integer" NovelPolymorphism="Free Text" FeatureNumber=”Number” />
<GenotypeAnnotation Annotation=”Structured Text”/>
<FeatureCoordinate FeatureNumber=”Number” FeatureStart=
Number” FeatureStop=”Number”/>
</Genotyping>
</Sample>
</Lab>
</IHIW_Report>
```

IHIWS XML

So far, most of vendors agreed on generating HML and XML output files

GOOD NEWS!

Although we have clearly defined HML schema,
There are so MANY VARIATIONS in HML output, especially reporting novel variants!

We would like to reach an agreement before the end of this summit!

Bad News

<glstring>HLA-A*01:01:01:01+HLA-A*29:02:01</glstring>

<reference-sequence id="ref8" name="HLA-A*29:02:01:01" />

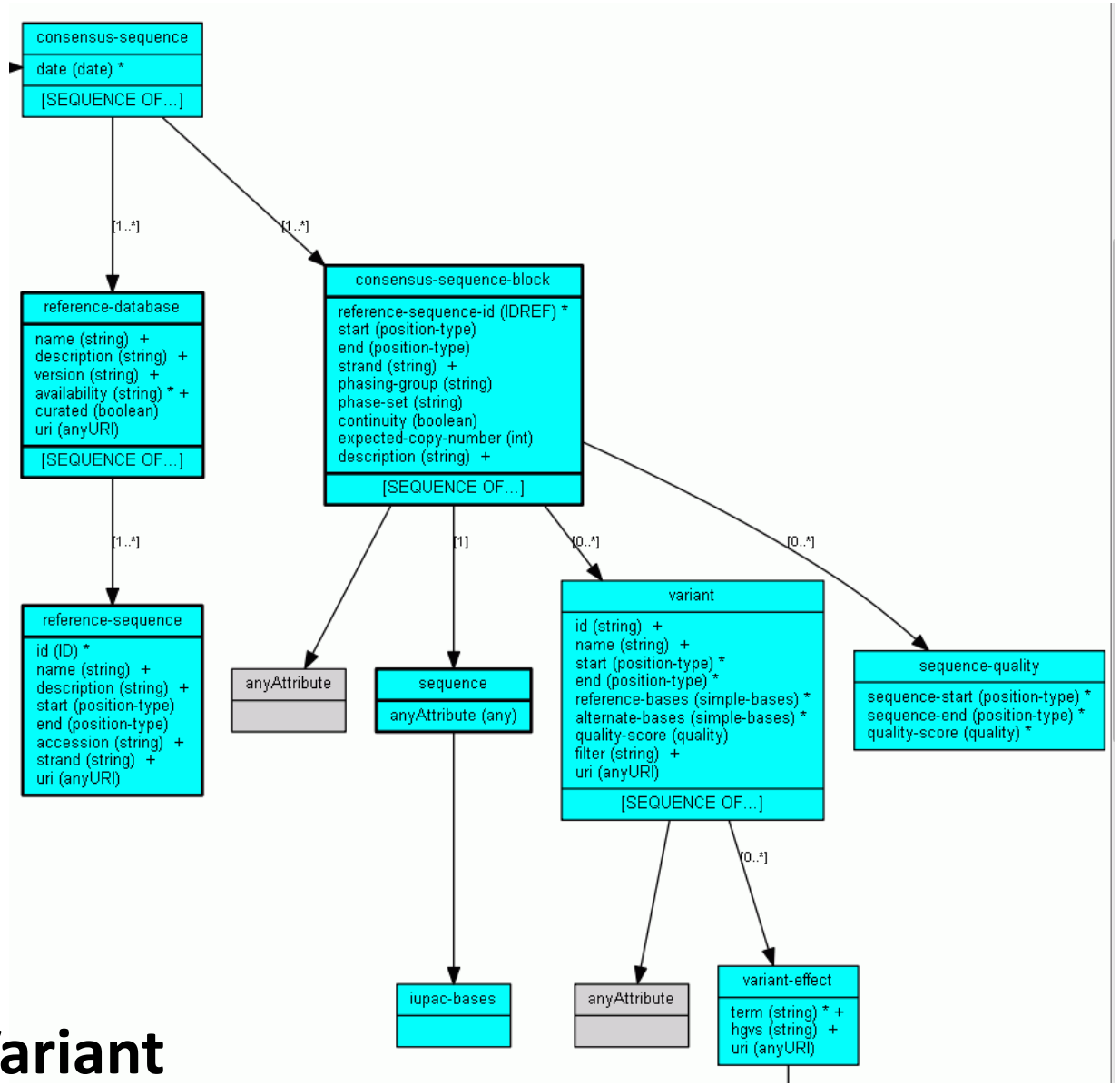
“Assume there is a mismatch in intron compared to HLA-A*29:02:01:01”
No “variant” information

<glstring>HLA-A*01:01:01:01+HLA-A*29:02:01:01</glstring>

<reference-sequence id="ref8" name="HLA-A*29:02:01:01" />

“Assume this is perfect match to HLA-A*29:02:01:01”
When we review the software,
There is a mismatch, but not reported in HML

A*02:01:01L
HLA- A*02:01:01”02L?



Variations in Variant

<reference-sequence id="ref2" name="HLA-A*68:01:02:01" start="0" end="3644" />

<consensus-sequence-block reference-sequence-id="ref2" start="0" end="3516" phase-set="1" continuity="true">

Actual length of consensus sequence = 3516!!!

Genomic Sequence Data for A*68:01:02:01 (HLA00116) says 3517 bps

<variant end="1770" start="1769" alternate-bases="T" reference-bases="C"></variant>

The character position 1770 of HLA00116 is “G”.

The character position 1770 of consensus sequence is “G”.

Problem seeing in reporting variants in HML!

Impression

HLA genes are so polymorphic, and so many Insertions/deletions
Vendors tend to make their own reference sequence and own base positions

Impossible to locate “position” of a specific nucleotide unless everyone agrees on standard

Question!

Do we need to call “Novel” polymorphism if some exon sequences are missing?
No intron sequences are available?
How do we report variants if gene sequence is not complete?
Which “reference” sequence should we use for missing sequence?
How do we define base positions?