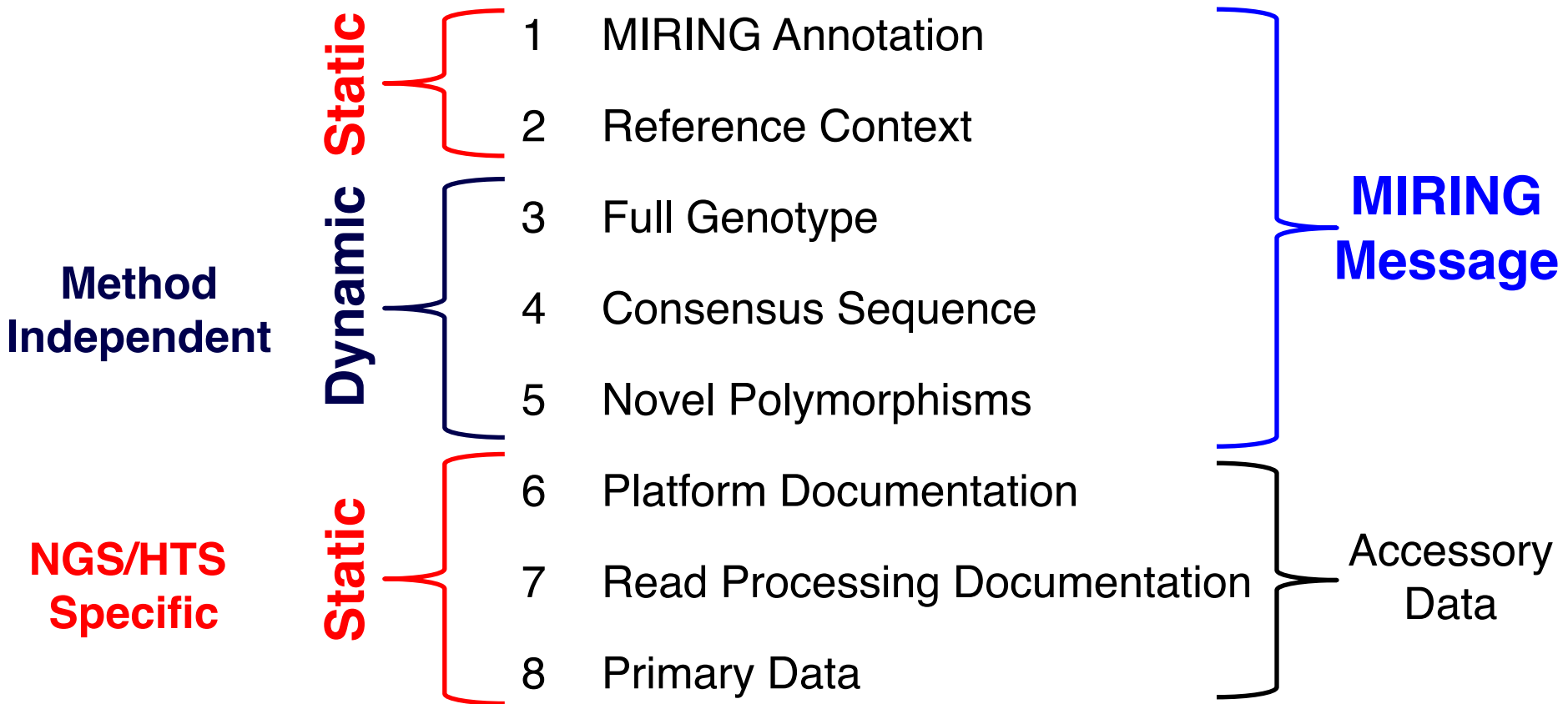# MIRING & HML Validation

Bob Milius, PhD

Bioinformatics Research

NMDP®/Be The Match®

Minimum **I**nformation for **R**eporting **I**mmunogenomic **N**GS **G**enotyping

Define an NGS Genotyping Result with 8 Categories of Information

Static
1   MIRING Annotation
2   Reference Context

Dynamic
3   Full Genotype
4   Consensus Sequence
5   Novel Polymorphisms

Static
6   Platform Documentation
7   Read Processing Documentation
8   Primary Data

Method Independent

NGS/HTS Specific

MIRING Message

Accessory Data

http://igdawg.org/miring.html

MIRING & HML

- MIRING is a set of guiding principles

- HML 1.0 is a technical specification that allows information conforming to MIRING principles to be exchanged
  - But HML doesn't have to be MIRING compliant
  - You can perfectly valid HML, yet doesn't contain much information.
  - XML Schema

Google "xml schema"

An **XML Schema** is a language for expressing constraints about **XML** documents. There are several different **schema** languages in widespread use, but the main ones are Document Type Definitions (DTDs), Relax-NG, Schematron and W3C **XSD** (**XML Schema** Definitions).

XML Schema - W3C

https://www.w3.org/standards/**xml**/**schema**
World Wide Web Consortium

Histoimmunogenetics Markup Language 1.0: Reporting next generation sequencing-based HLA and KIR genotyping

CrossMark

Robert P. Milius [a,*], Michael Heuer [a], Daniel Valiga [a], Kathryn J. Doroschak [a], Caleb J. Kennedy [a], Yung-Tsi Bolon [a], Joel Schneider [a], Jane Pollack [a], Hwa Ran Kim [b], Nezih Cereb [b], Jill A. Hollenbach [c], Steven J. Mack [d], Martin Maiers [a]

[a] National Marrow Donor Program, MN, USA
[b] HistoGenetics LLC, USA
[c] University of California – San Francisco, USA
[d] Children's Hospital & Research Center Oakland, Oakland, CA, USA

ARTICLE INFO

ABSTRACT

We present an electronic format for exchanging data for HLA and KIR genotyping with extensions for next-generation sequencing (NGS). This format addresses NGS data exchange by refining the Histoimmunogenetics Markup Language (HML) to conform to the proposed Minimum Information for Reporting Immunogenomic NGS Genotyping (MIRING) reporting guidelines (miring.immunogenomics.org). Our refinements of HML include two major additions. First, NGS is supported by new XML structures to capture additional NGS data and metadata required to produce a genotyping result, including analysis-dependent (dynamic) and method-dependent (static) components. A full genotype, consensus sequence, and the surrounding metadata are included directly, while the raw sequence reads and platform documentation are externally referenced. Second, genotype ambiguity is fully represented by integrating Genotype List Strings, which use a hierarchical set of delimiters to represent allele and genotype ambiguity in a complete and accurate fashion. HML also continues to enable the transmission of legacy methods (e.g. site-specific oligonucleotide, sequence-specific priming, and Sequence Based Typing (SBT)), adding features such as allowing multiple group-specific sequencing primers, and fully leveraging techniques that combine multiple methods to obtain a single result, such as SBT integrated with NGS.

5

BE THE MATCH®

My Account   Physicians   News

For more on HML, see the Bioinformatics HML information page.

Public XML Schemas

Schemas on this page can be referenced either using the "http" OR "https" protocols. Some software may be unable to negotiate a secure connection if using "https" so be aware of your audience. All *submissions* of HML data should be sent over a secure transport layer.

- ○ HML 0.2 (diagram) **

- ○ HML 0.3 (diagram) **

- ○ HML 0.3.3 (diagram) **

- ○ HML 1.0 (diagram)

- ○ **HML 1.0.1** (diagram) **

```
▼<!--


        Histoimmunogenetics Markup Language (HML)
        Copyright (c) 2014 National Marrow Donor Program (NMDP)

        This specification is free software; you can redistribute it and/or modify it
        under the terms of the GNU Lesser General Public License as published
        by the Free Software Foundation; either version 3 of the License, or (at
        your option) any later version.

        This specification is distributed in the hope that it will be useful, but WITHOUT
        ANY WARRANTY; with out even the implied warranty of MERCHANTABILITY or
        FITNESS FOR A PARTICULAR PURPOSE.  See the GNU Lesser General Public
        License for more details.

        You should have received a copy of the GNU Lesser General Public License
        along with this specification;  if not, write to the Free Software Foundation,
        Inc., 59 Temple Place, Suite 330, Boston, MA 02111-1307  USA.

        > http://www.gnu.org/licenses/lgpl.html

  -->
▼<xs:schema xmlns="http://schemas.nmdp.org/spec/hml/1.0.1"
  xmlns:hmlns="http://schemas.nmdp.org/spec/hml/1.0.1" xmlns:xs="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified" attributeFormDefault="unqualified"
  targetNamespace="http://schemas.nmdp.org/spec/hml/1.0.1" version="1.0.1">
    <!--  HML document root  -->
  ▼<xs:element name="hml" id="hml" nillable="false">
    ▼<xs:annotation>
      ▼<xs:documentation>
          Root element of the document identifying it as an HML message. Must contain the version of HML that
          the modeled data in this document uses. Children: --------- - property (optional, qty: 0 or more) -
          hmlid (optional, qty: 1) - reporting-center (optional, qty: 1) - Required for NMDP samples - sample
          (required, qty: 1 or more) - typing-test-names (optional, qty: 0 or more) Attributes: ----------- -
          version: Version of HML the document follows (required) - project-name: Name of the typing project
          (optional)
      </xs:documentation>
    </xs:annotation>
    ▼<xs:complexType>
      ▼<xs:sequence>
        <xs:element name="property" type="hmlns:property" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element name="hmlid" type="hmlns:hmlid" minOccurs="0" maxOccurs="1"/>
        <xs:element ref="reporting-center" minOccurs="0" maxOccurs="1"/>
```

BeTheMatch Schema Reposit ×

schemas.nmdp.org

BE THE MATCH®

Give   Join   Volunteer   Shop

My Account   Physicians   News

For more on HML, see the Bioinformatics HML information page.

Public XML Schemas

Schemas on this page can be referenced either using the "http" OR "https" protocols. Some software may be unable to negotiate a secure connection if using "https" so be aware of your audience. All *submissions* of HML data should be sent over a secure transport layer.
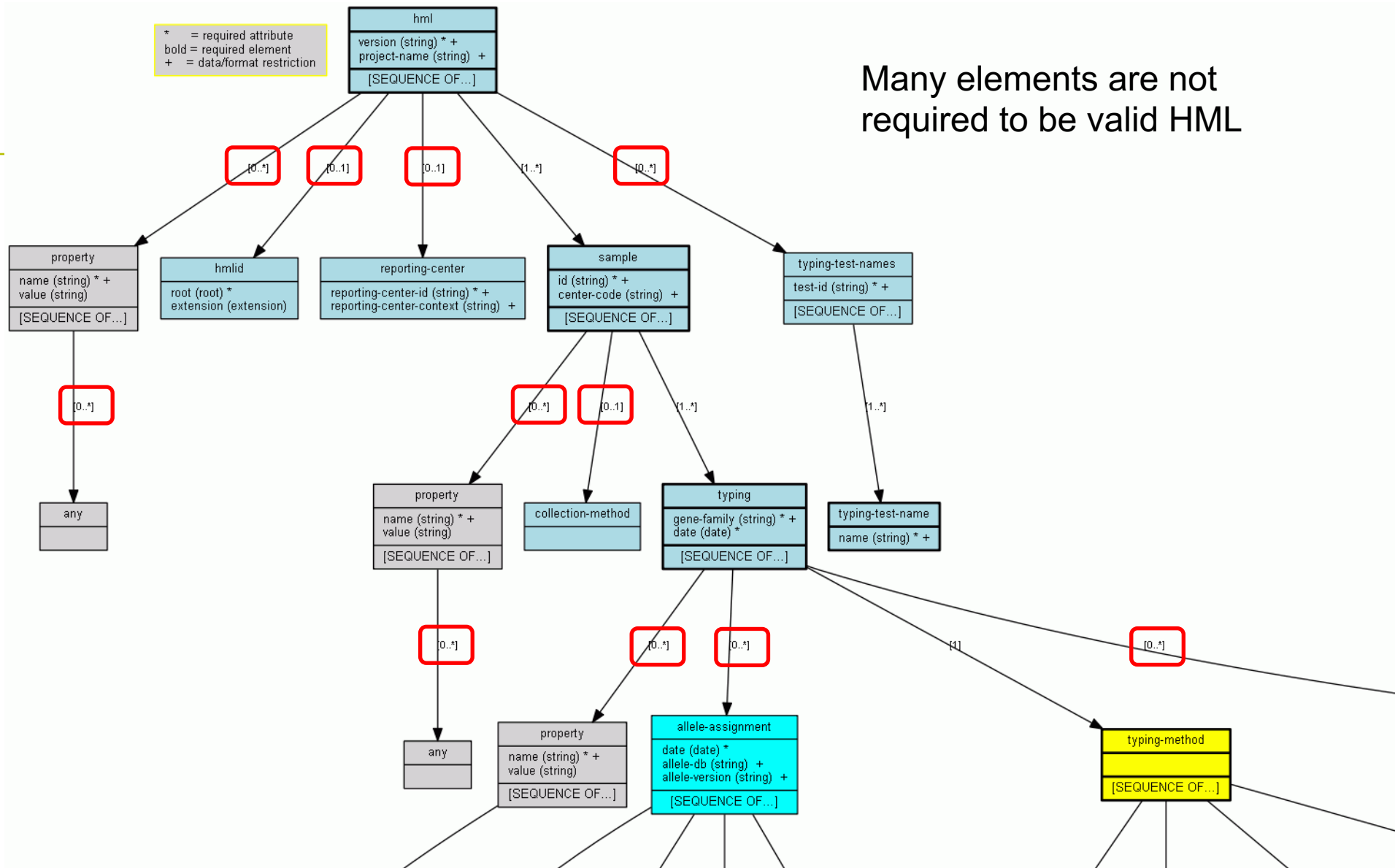
- ○ HML 0.2 (diagram) **

- ○ HML 0.3 (diagram) **

- ○ HML 0.3.3 (diagram) **

- ○ HML 1.0 (diagram)

- ○ **HML 1.0.1** (diagram) **
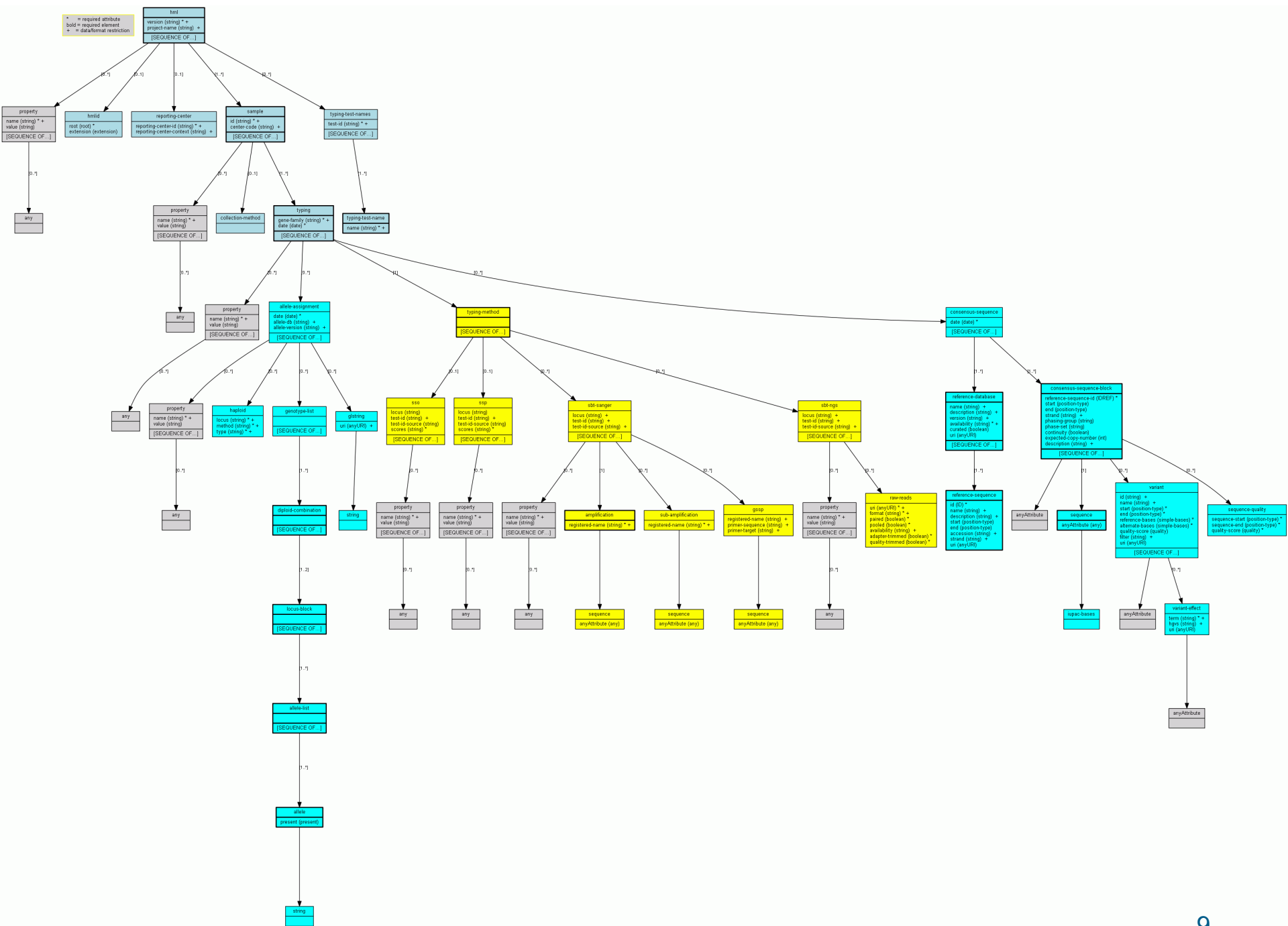
Many elements are not required to be valid HML

Tiny Valid HML 1.0.1

```xml
<?xml version="1.0" ?>
<hml xmlns="http://schemas.nmdp.org/spec/hml/1.0.1"
    version="1.0.1" >
    <sample id="1">
        <typing gene-family="HLA" date="2016-07-01">
            <typing-method/>
        </typing>
    </sample>
</hml>
```

Command line validation

```
$ xmllint --noout --schema hml-1.0.1.xsd tiny.hml

tiny.hml validates

$
```

MIRING Validator

- First, is it valid HML?
  - Is it well formed?
  - XML Schema

- Second, does it contain elements necessary to support MIRING?
  - Schematron

Google "xml schema"

An **XML Schema** is a language for expressing constraints about **XML** documents. There are several different **schema** languages in widespread use, but the main ones are Document Type Definitions (DTDs), Relax-NG, **Schematron** and W3C **XSD** (**XML Schema** Definitions).

XML Schema - W3C

https://www.w3.org/standards/**xml**/**schema**
World Wide Web Consortium

XML Schema vs Schematron

- XML Schema
  - Grammar based
    - e.g., cardinality, data types, value sets
  - XSD
  - W3C

- Schematron
  - Rules based on patterns
    - Assertions applied to a specific context within the document
    - e.g., "start" must be smaller than "end"
  - XSLT and XPath
  - ISO/IEC

Tiers of Validation

- **Tier 1**
  - XML Schema to validate both MIRING and HML schemas. We can check simple MIRING and HML syntax rules, for existence of nodes or attributes.

- **Tier 2**
  - Use Schematron to validate some more complicated rules. More complicated data comparisons (Start < End) can be done here. Schematron is an XML validation rules engine, so we use it to do comparisons, etc. that XML Schema validation can't do.

- **Tier 3**
  - Tier 3 is where we assigned external lookups, such as Method Registries, Reference Sequence Databases, GLString validations, etc. This is outside of scope for now.

HOW TO READ A VALIDATION LINE

Example of a validation line

```
hml-result severity="fatal"
description[306,138] Value '' with length = '0' is not facet-valid with respect to minLength '5' for type '#AnonType_uriraw-reads'. /description
solution Verify that your HML file is well formed, and conforms to http://schemas.nmdp.org/spec/hml/1.0.1/hml-1.0.1.xsd /solution /hml-result
```

`hml-result severity="fatal"` This line shows the hml or miring result which is split into two sections. The severity levels allows for the validator to know if it the issue should be allowed or not this can either be fatal, hml, miring, warning, or info. HML,miring, and fatal errors will stop the validation and must be fixed immeditatly. For miring results, you will see a `miring-rule-id` which specifies which rule is being violated.

`[306,138]` Represents the row and column number respectfully (Note: If you are missing a matching end tag the system might report the location to be the end of the file. Please just refer to the description if this happens)

`Value '' with length = '0' is not facet-valid with respect to minLength '5' for type '#AnonType_uriraw-reads'.` is an autogenerated error message based from SAXParser Schema Error codes this will always appear in the description nodes

`Verify that your HML file is well formed, and conforms to http://schemas.nmdp.org/spec/hml/1.0.1/hml-1.0.1.xsd` is a statement made by the developers on a start on how to fix this specific issue and will always be in the solution node

On occasions you will see a location, this will be an XPath leading to the error, I suggest reading this webpage on how to read it

RULES

Miring Element	Miring Description	Rule ID	Severity	Rule Description
1	Message Annotation			
1.1	Unique Miring Message Identifier	1.1.a	miring	Document has exactly one hmlid node directly under the root hml node
		1.1.c	info	if hmlid is an OID the hmlid node should have a root and an extension attribute.
1.2	Message Generator Contact Information	1.2.a	miring	Document has exactly one reporting-center node directly under the root hml node
		1.2.b	warning	reporting-center has one reporting-center-id and one reporting-center-context attributes
1.3	Platform Documentation (MIRING element 6) Reference	1.3.a	miring	Every sso ssp sbt-sanger sbt-ngs node should have a test-id and test-id-source attributes
		1.3.b	info	test-id and test-id-source are in a recognized format. NCBI-GTR
1.4	Read Processing Documentation (MIRING element 7) Reference			

Be The Match
Validator code is on GITHUB

Table 1
Mapping of MIRING elements to HML solutions.

MIRING elements		HML solution example with comment
1	Message Annotation	
1.1	Unique MIRING Message Identifier	`<hmlid root= "x.x.x.x" extension= "x" />` root may be ISO Object Identifier (OID) or Universally Unique Identifier (UUID)

1.2	Message Generator Contact Information	Using an NCBI GTR Lab ID:
		`<reporting-center`
		`  reporting-center-id= "000000"`
		`  reporting-center-context="NCBI-GTR" />`

1.3	Platform Documentation	Using an NCBI GTR Test ID:
		`<sbt-ngs ...`
		`  test-id="GTR000000000.0"`
		`  test-id-source="NCBI-GTR" />`

1.4	Read Processing Documentation Reference	Using an NCBI GTR Test ID:
		`<sbt-ngs ...`
		`  test-id="GTR000000000.0"`
		`  test-id-source="NCBI-GTR" />`

1.5	Primary Data Availability	`<raw-reads ... availability="true" ... />`
1.6	Primary Data Reference	`<raw-reads ...` `  uri="http://www.ncbi.nlm.nih.gov/sra/SRX000000" ... />`
2	Reference Context	

2.1	Reference Sequence Database Version for Allele Calling	For allele assignment, the allele database and version is captured in attributes:
		`<allele-assignment ...`
		`  allele-db="IMGT-HLADB"`
		`  allele-version="3.19.0" >`
		`    <glstring>...</glstring>`
		`</allele-assignment>`
		For sequence variant, the reference database and sequence is identified through the reference-database and reference-sequence elements.
		`<reference-database >`
		`  <reference-sequence ... />`
		`  <reference-database... />`
		`</reference-database>`

2.2	Individual Reference Sequences Applied	`<reference-sequence`
		`  id="0"`
		`  name="HSCHR6_MHC_MCF_CTG1"`
		`  start="0"`
		`  end="4827813"`
		`  accession="CCDS34373.12"`
		`  uri="http://www.ncbi.nlm.nih.gov/nuccore/GL000254.2"/>`

2.2.1	Reference Sequence Identifier	`<reference-sequence ... id="0" ... />`

2.3	Reference Sequence Source Type	`<reference-database ...`
		`  public="true"`
		`  curated="true"`
		`  />`

Things to remember…

- Just because your HML passes both HML and MIRING validation doesn't mean it's okay

- The validator doesn't check…
  - if your coordinates are 0-base
  - if your reference URIs and IDs match with the coordinates you supply
  - if the URIs you enter actually resolve to anything real
  - if the strand you enter is correct for the reference you supply
  - If the variant coordinates are within the window of the reference sequence that makes up your consensus-sequence-block
  - etc

- The creator of the HML is still responsible for its content, that it's complete and makes sense

Links

- Web GUI
  - http://miring.b12x.org
- Source code
  - https://github.com/nmdp-bioinformatics/MiringValidator

- HML manuscript
  - http://www.ncbi.nlm.nih.gov/pubmed/26319908
- MIRING manuscript
  - http://www.ncbi.nlm.nih.gov/pubmed/26407912

Acknowledgements

- NMDP Bioinformatics
  - Ben Matern
  - Anubhav Balakrishnan
  - Jason Brelsford
  - Dan Valiga
  - Clement Ellango
  - Mike George
  - Yung Tsi Bolon
  - Jane Pollack
  - Bob Milius
  - Martin Maiers

- CHORI
  - Steven J. Mack
- Stanford
  - Kazutoyo Osoegawa
  - Chia-Jung Chang
  - Marcelo Fernandez-Vina
- *(please let know who else…)*

Thank you!

Questions?