CINECA

Common Infrastructure for National Cohorts in Europe, Canada, and Africa

Beacon v2 Model
Data Ingestion &BFF

CRG Centre for Genomic Regulation

This project has received funding from the European Union's Horizon 2020 research and Innovation programme under grant agreement No. 825775

This project has received funding from the Canadian Institute of Health Research under grant agreement #404896

CIHR IRSC

Beacon v2 is a protocol / specification

```
{
  "$schema": "http://json-schema.org/draft-07/schema",
  "title": "Beacon Configuration",
  "description": "Files complaint with this schema are the configuration ones. The details re
  "type": "object",
  "properties": {
    "$schema": {
      "$ref": "../common/beaconCommonComponents.json#/definitions/$schema"
    },
    "maturityAttributes": {
      "description": "Declares the level of maturity of the Beacon instance.",
      "type": "object",
      "properties": {
        "productionStatus": {
          "description": "`DEV`= 'Service potentially unstable, not real data', which availab
          "type": "string",
          "enum": ["DEV", "TEST", "PROD"]
        }
      }
    },
    "securityAttributes": {
      "description": "Configuration of the security aspects of the Beacon. By default, a Beac
      "type": "object",
      "properties": {
        "defaultGranularity": {
          "description": "Default granularity. Some responses could return higher detail, but
          "type": "string",
          "enum": ["boolean", "count", "aggregated", "record"],
```

Docs » Schemas (Draft v4) » Genomic Variations

Field	description	type	properties
alternateBases	Alternate bases for this variant (starting from `start`). *Accepted values: IUPAC codes for nucleotides (e.g. 'https://www.bioinformatics.org/sms/iupac.html'). N is a wildcard, that denotes the position of any base, and can beused as a standalone base of any type or within a partially knownsequence. As example, a query of* `ANNT` *the Ns can take take any form of[ACGT] and will match* `ANNT`, `ACNT`, `ACCT`, `ACGT` *... and so forth.* an *empty value* is used in the case of deletions with the maximally trimmed, deleted sequence being indicated in `ReferenceBases` *Categorical variant queries, e.g. such* not *being represented through sequence & position, make use of the* `variantType` *parameter.* either `alternateBases` or `variantType` is required.	string	NA
caseLevelData		array	alleleOrigin, analysisId, biosampleId, clinicalInterp individualId, phenotypicF zigosity
frequencyInPopulations	NA	array	frequencies, sourceRefer version
identifiers	NA	NA	NA
molecularAttributes	NA	NA	NA
position	This section groups all attributes that allows to 'identify' a variant via its position in the genome.	object	assemblyId, refseqId, sta

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

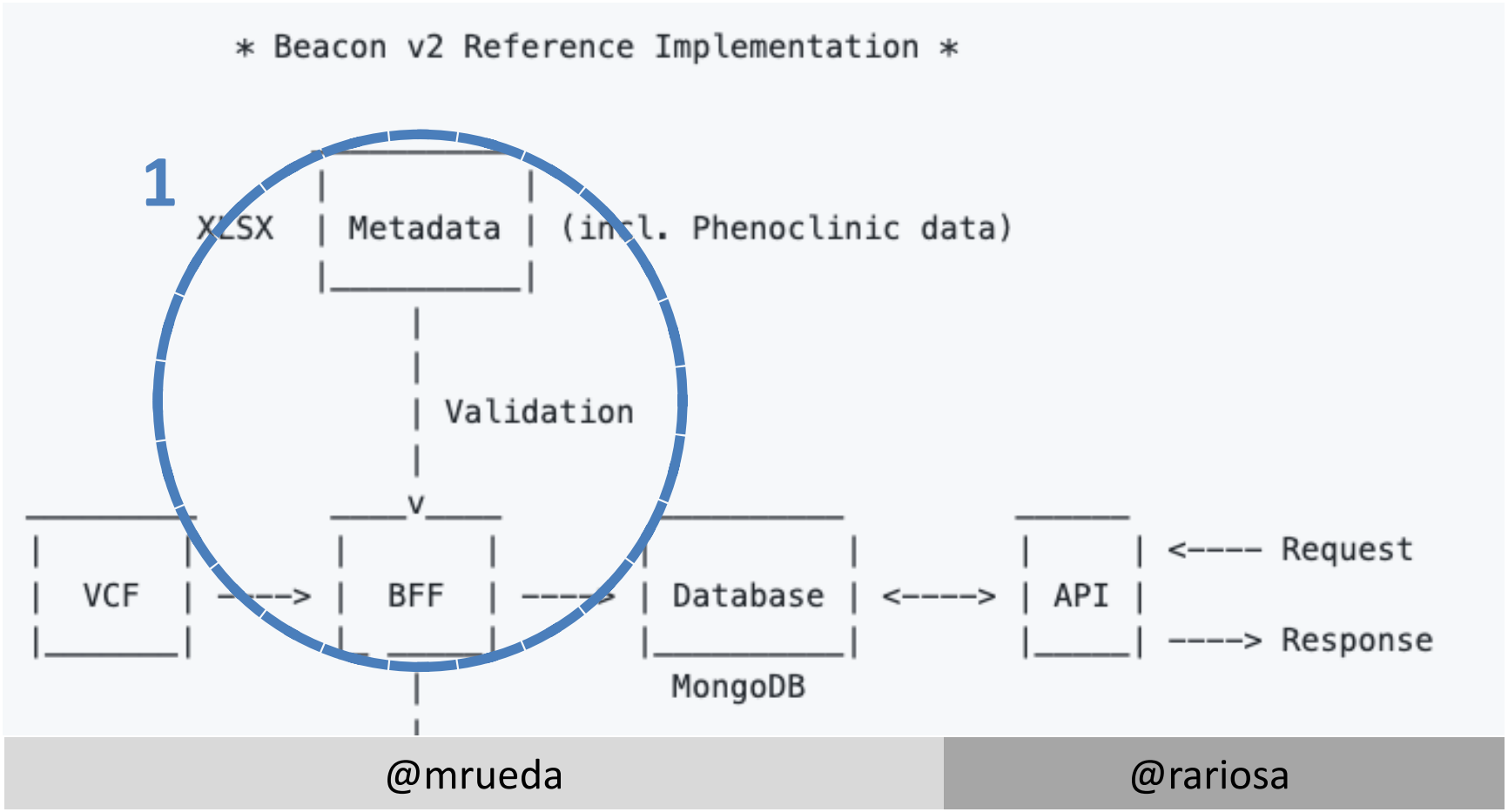
CINECA

Data ingestion consists of 3 steps

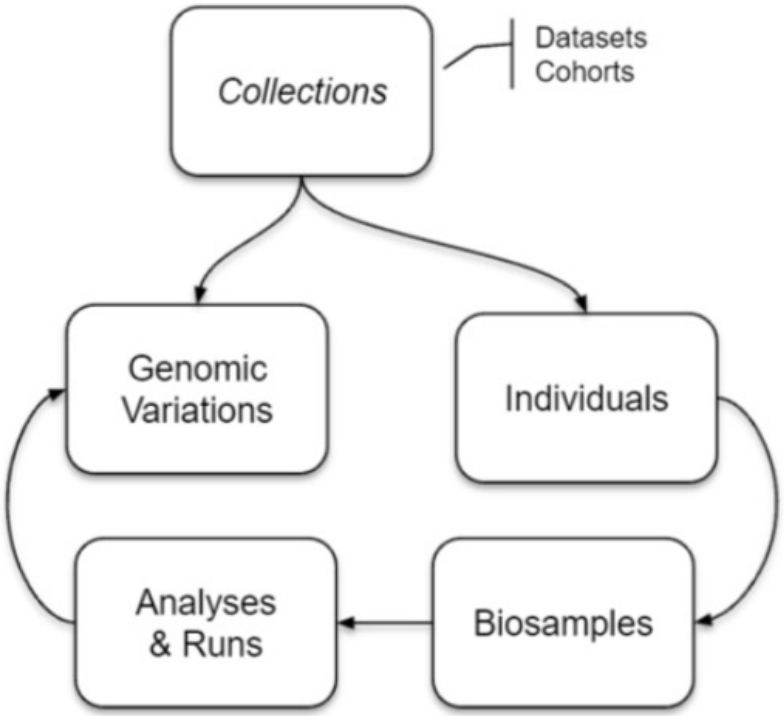
1. Transforming metadata
2. Transforming genomic variations (VCF)
3. Load data into MongoDB

Data ingestion consists of 3 steps

1. Transforming metadata



The idea is to transform your data to the Beacon v2 Models



Docs » Schemas (Draft v4) » Genomic Variations

Field	description	type	properties
alternateBases	Alternate bases for this variant (starting from `start`). *Accepted values: IUPAC codes for nucleotides (e.g. 'https://www.bioinformatics.org/sms/iupac.html'). N is a wildcard, that denotes the position of any base, and can beused as a standalone base of any type or within a partially knownsequence. As example, a query of* `ANNT` *the Ns can take take any form of[ACGT] and will match* `ANNT`, `ACNT`, `ACCT`, `ACGT` *... and so forth.* an *empty value* is used in the case of deletions with the maximally trimmed, deleted sequence being indicated in `ReferenceBases` *Categorical variant queries, e.g. such* not *being represented through sequence & position, make use of the* `variantType` *parameter.* either `alternateBases` or `variantType` is required.	string	NA
caseLevelData		array	alleleOrigin, analysisId, biosampleId, clinicalInter individualId, phenotypicE zigosity
frequencyInPopulations	NA	array	frequencies, sourceRefer version
identifiers	NA	NA	NA
molecularAttributes	NA	NA	NA
position	This section groups all attributes that allows to 'identify' a variant via its position in the genome.	object	assemblyId, refseqId, sta

7 Entities

Metadata follows Bv2 Models

Input: Excel file with 6 sheets (+1 for g_variations)

	A	B	C	D	E	F	G	H	I
5	bwa-0.7.8	2021-12-29	synthetic_sample_4	UK1_4	HG00100	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
6	bwa-0.7.8	2021-12-29	synthetic_sample_5	UK1_5	HG00101	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
7	bwa-0.7.8	2021-12-29	synthetic_sample_6	UK1_6	HG00102	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
8	bwa-0.7.8	2021-12-29	synthetic_sample_7	UK1_7	HG00103	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
9	bwa-0.7.8	2021-12-29	synthetic_sample_8	UK1_8	HG00105	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
10	bwa-0.7.8	2021-12-29	synthetic_sample_9	UK1_9	HG00106	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
11	bwa-0.7.8	2021-12-29	synthetic_sample_10	UK1_10	HG00107	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
12	bwa-0.7.8	2021-12-29	synthetic_sample_11	UK1_11	HG00108	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
13	bwa-0.7.8	2021-12-29	synthetic_sample_12	UK1_12	HG00109	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
14	bwa-0.7.8	2021-12-29	synthetic_sample_13	UK1_13	HG00110	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
15	bwa-0.7.8	2021-12-29	synthetic_sample_14	UK1_14	HG00111	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
16	bwa-0.7.8	2021-12-29	synthetic_sample_15	UK1_15	HG00112	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
17	bwa-0.7.8	2021-12-29	synthetic_sample_16	UK1_16	HG00113	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
18	bwa-0.7.8	2021-12-29	synthetic_sample_17	UK1_17	HG00114	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
19	bwa-0.7.8	2021-12-29	synthetic_sample_18	UK1_18	HG00115	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
20	bwa-0.7.8	2021-12-29	synthetic_sample_19	UK1_19	HG00116	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
21	bwa-0.7.8	2021-12-29	synthetic_sample_20	UK1_20	HG00117	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
22	bwa-0.7.8	2021-12-29	synthetic_sample_21	UK1_21	HG00118	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
23	bwa-0.7.8	2021-12-29	synthetic_sample_22	UK1_22	HG00119	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
24	bwa-0.7.8	2021-12-29	synthetic_sample_23	UK1_23	HG00120	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
25	bwa-0.7.8	2021-12-29	synthetic_sample_24	UK1_24	HG00121	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
26	bwa-0.7.8	2021-12-29	synthetic_sample_25	UK1_25	HG00122	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0

analyses | biosamples | cohorts | datasets | genomicVariations | individuals | runs | +

Possible scenario: SQL partial dump/export

Metadata are transformed to BFF

Output = Beacon Friendly Format (BFF)

7 JSON files (JSON array)

JSON → Documents (Objects)
Key: Value

analyses.json


```
[
  {
    "aligner" : "bwa-0.7.8",
    "analysisDate" : "2021-12-29",
    "biosampleId" : "synthetic_sample_1",
    "id" : "UK1_1",
    "individualId" : "HG00096",
    "pipelineName" : "Pipeline-panel-0001-v1",
    "pipelineRef" : "https://doi.org/10.48511/workflowhub.workflow.111.1",
    "runId" : "SRR10903401",
    "variantCaller" : "GATK4.0"
  },
  {
    "aligner" : "bwa-0.7.8",
    "analysisDate" : "2021-12-29",
    "biosampleId" : "synthetic_sample_2",
    "id" : "UK1_2",
    "individualId" : "HG00097",
    "pipelineName" : "Pipeline-panel-0001-v1",
    "pipelineRef" : "https://doi.org/10.48511/workflowhub.workflow.111.1",
    "runId" : "SRR10903401",
    "variantCaller" : "GATK4.0"
  },
  {
    "aligner" : "bwa-0.7.8",
    "analysisDate" : "2021-12-29",
    "biosampleId" : "synthetic_sample_3",
    "id" : "UK1_3",
    "individualId" : "HG00099",
    "pipelineName" : "Pipeline-panel-0001-v1",
    "pipelineRef" : "https://doi.org/10.48511/workflowhub.workflow.111.1",
    "runId" : "SRR10903401",
    "variantCaller" : "GATK4.0"
  },
```

Metadata transformation is crucial

- **Cons ☹**
  - Users must map their terms to Bv2 *default schemas* to enable DB⇠⇢API queries (no need to share everything!!)
  - Deeply nested structures – bumps expected
  - Suboptimal option if you already have data structured according to *OMOP CDM* , *FHIR*, etc.
- **Pros ☺**
  - **Consistency** (From DB → Responses)
  - Metadata are **validated** against Bv2 models schemas (e.g., string, date, array, regex, etc.)
  - **Documentation** available (RTD)
  - You can skip EXCEL → map terms → **JSON** (if you know how)

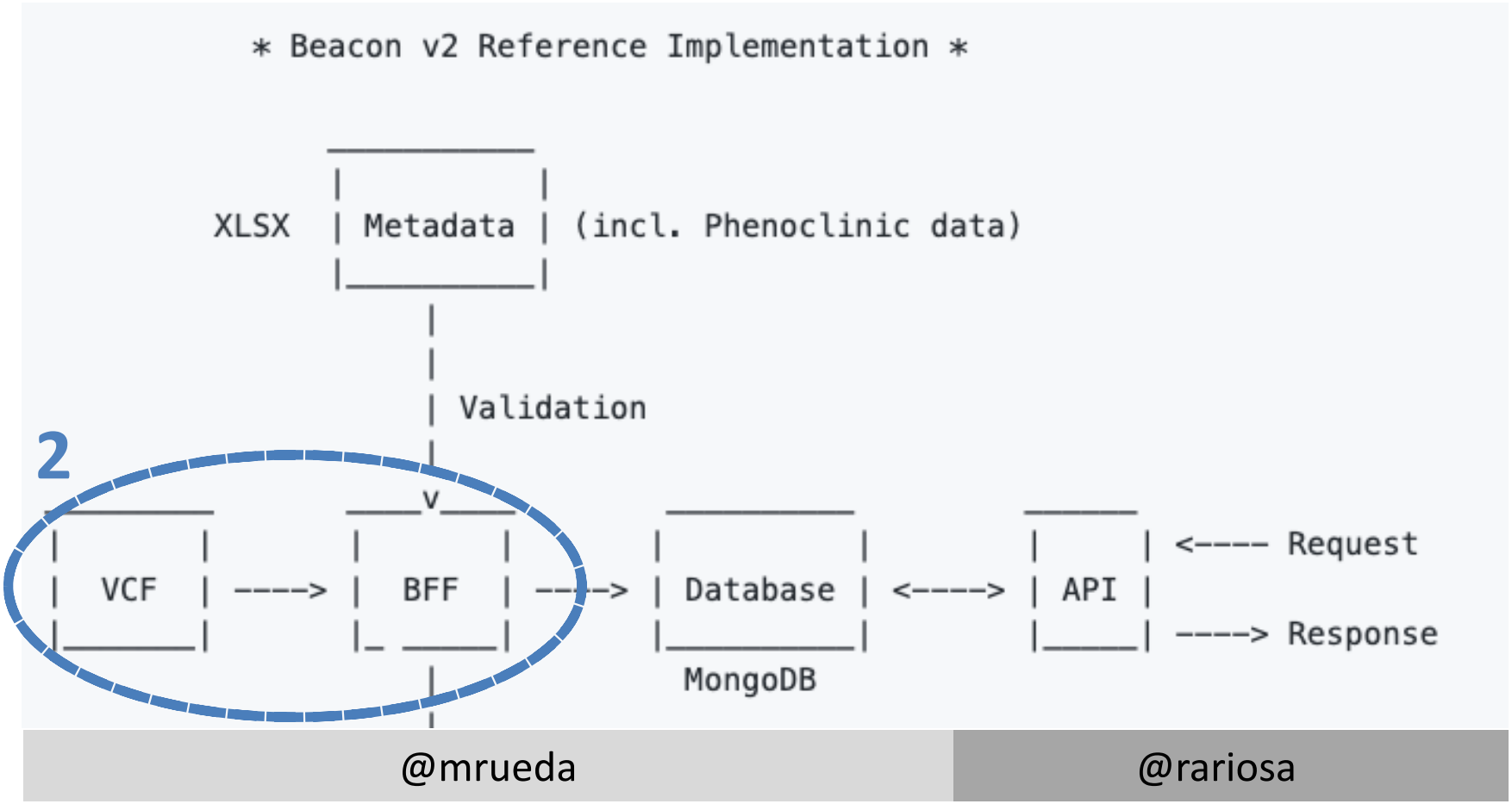
B2RI: Metadata → BFF

```
./bff-validator -i Beacon-v2-Models_template.xlsx --out-dir bff
```

```
****************************************
*  Beacon v2 Reference Implementation  *
* - BEACON FRIENDLY FORMAT VALIDATOR - *
*           Version: 2.0.0             *
*   (C) 2021-2022 Manuel Rueda, PhD    *
*    GNU General Public License v3     *
****************************************

==== SCHEMA: ANALYSES ====
Hurray! No errors found
Printing <bff/analyses.json> file

==== SCHEMA: BIOSAMPLES ====
Hurray! No errors found
Printing <bff/biosamples.json> file

==== SCHEMA: COHORTS ====
Hurray! No errors found
Printing <bff/cohorts.json> file

==== SCHEMA: DATASETS ====
Hurray! No errors found
Printing <bff/datasets.json> file

==== SCHEMA: INDIVIDUALS ====
Hurray! No errors found
Printing <bff/individuals.json> file

==== SCHEMA: RUNS ====
Hurray! No errors found
Printing <bff/runs.json> file
```

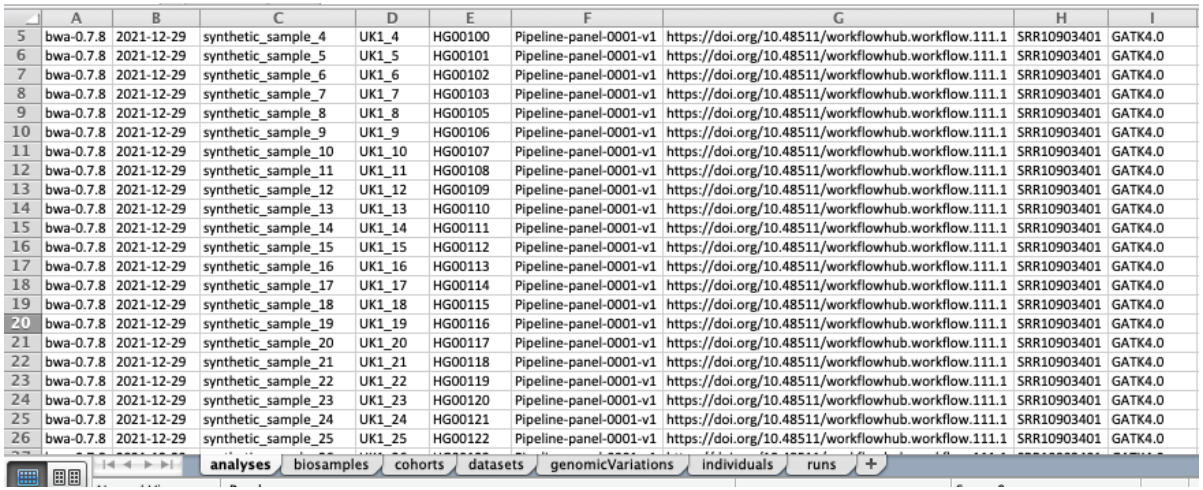
Data ingestion consists of 3 steps

2. Transforming genomic variations



Transforming genomic variations

1. Use XLSX template



	A	B	C	D	E	F	G	H	I
5	bwa-0.7.8	2021-12-29	synthetic_sample_4	UK1_4	HG00100	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
6	bwa-0.7.8	2021-12-29	synthetic_sample_5	UK1_5	HG00101	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
7	bwa-0.7.8	2021-12-29	synthetic_sample_6	UK1_6	HG00102	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
8	bwa-0.7.8	2021-12-29	synthetic_sample_7	UK1_7	HG00103	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
9	bwa-0.7.8	2021-12-29	synthetic_sample_8	UK1_8	HG00105	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
10	bwa-0.7.8	2021-12-29	synthetic_sample_9	UK1_9	HG00106	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
11	bwa-0.7.8	2021-12-29	synthetic_sample_10	UK1_10	HG00107	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
12	bwa-0.7.8	2021-12-29	synthetic_sample_11	UK1_11	HG00108	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
13	bwa-0.7.8	2021-12-29	synthetic_sample_12	UK1_12	HG00109	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
14	bwa-0.7.8	2021-12-29	synthetic_sample_13	UK1_13	HG00110	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
15	bwa-0.7.8	2021-12-29	synthetic_sample_14	UK1_14	HG00111	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
16	bwa-0.7.8	2021-12-29	synthetic_sample_15	UK1_15	HG00112	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
17	bwa-0.7.8	2021-12-29	synthetic_sample_16	UK1_16	HG00113	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
18	bwa-0.7.8	2021-12-29	synthetic_sample_17	UK1_17	HG00114	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
19	bwa-0.7.8	2021-12-29	synthetic_sample_18	UK1_18	HG00115	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
20	bwa-0.7.8	2021-12-29	synthetic_sample_19	UK1_19	HG00116	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
21	bwa-0.7.8	2021-12-29	synthetic_sample_20	UK1_20	HG00117	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
22	bwa-0.7.8	2021-12-29	synthetic_sample_21	UK1_21	HG00118	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
23	bwa-0.7.8	2021-12-29	synthetic_sample_22	UK1_22	HG00119	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
24	bwa-0.7.8	2021-12-29	synthetic_sample_23	UK1_23	HG00120	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
25	bwa-0.7.8	2021-12-29	synthetic_sample_24	UK1_24	HG00121	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0
26	bwa-0.7.8	2021-12-29	synthetic_sample_25	UK1_25	HG00122	Pipeline-panel-0001-v1	https://doi.org/10.48511/workflowhub.workflow.111.1	SRR10903401	GATK4.0

analyses | biosamples | cohorts | datasets | genomicVariations | individuals | runs | +

2. Use B2RI's VCF data ingestion tool

How it's made?

- Input VCF file can be:
  - Version ≥ 4.1, Single/multi-sample (no gVCF)
  - Ref: hs37, hg19/38
  - Splits multiallelic variants to biallelic
- Re-Annotated with **SnpEff**
  - Remember: The objective is **SHARING**!!

  (Your own annotations → map terms → Excel / JSON)
- Serialize to JSON (genomicVariationsVcf.json.gz)

(Speed = 1M variants ~ 15 min * 1 core)

B2RI: VCF → BFF

```
./beacon vcf -i test.vcf.gz -p parameters.yaml
```

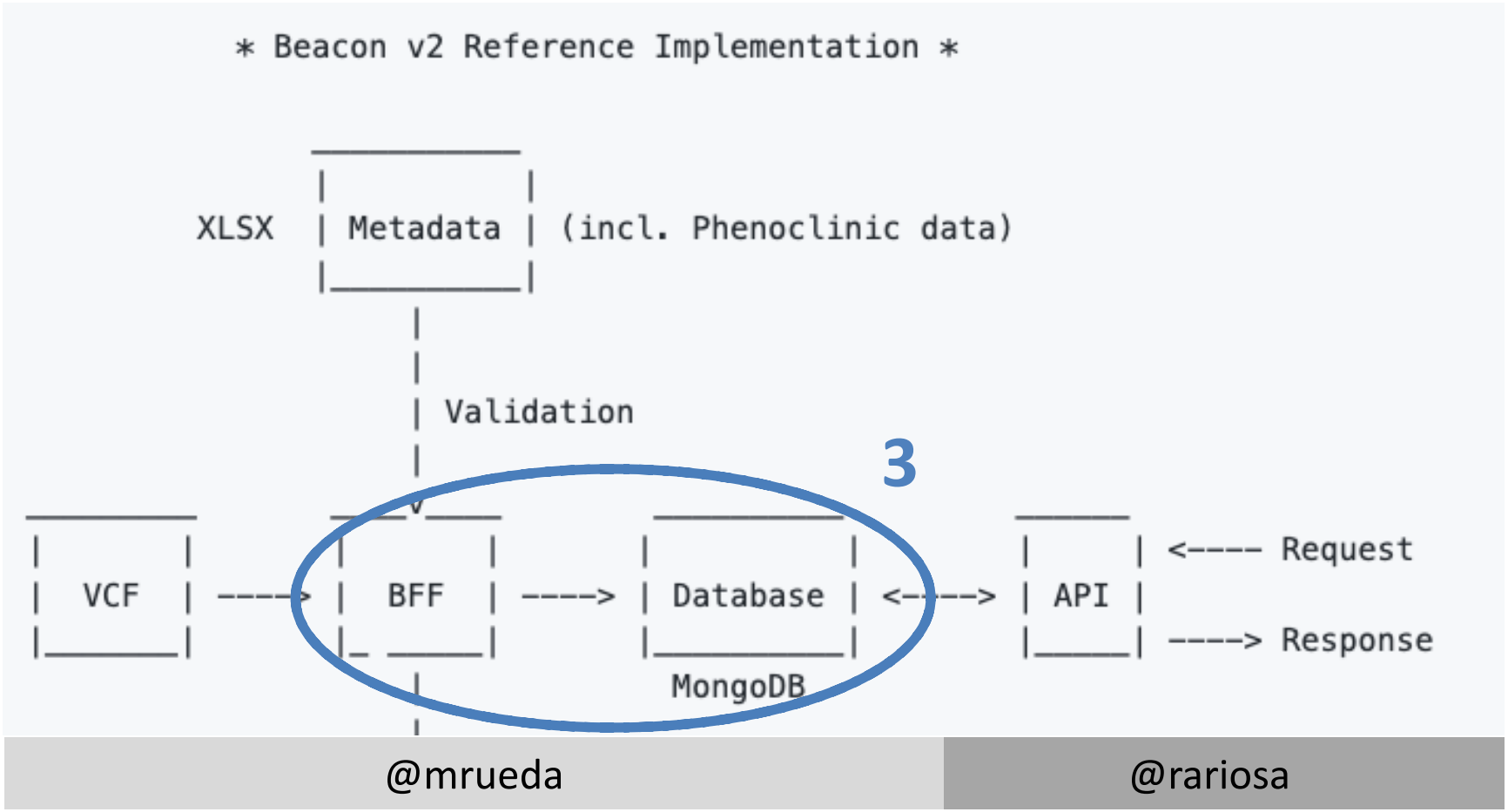
```
Info:
Info: Beacon 2.0.0
Info: Beacon exe: /pro/beacon-2.0.0/beacon
Info: (C) 2021 Manuel Rueda, PhD
Info: GNU General Public License v3
Info:
Info: ARGUMENTS USED:
Info: mode: vcf
Info: --i test.vcf.gz
Info: --p parameters.yaml
Info:
Info: CONFIGURATION VALUES:
Info: bash4bff          =>  /pro/beacon-2.0.0/BEACON/bin/run_vcf2bff.sh
Info: bash4html         =>  /pro/beacon-2.0.0/BEACON/bin/run_bff2html.sh
Info: bash4mongodb      =>  /pro/beacon-2.0.0/BEACON/bin/run_bff2mongodb.sh
Info: bcftools          =>  /pro/NGSutils/bcftools-1.11/bcftools
Info: bff2json          =>  /pro/beacon-2.0.0/BEACON/bin/bff2json.pl
Info: bff2mongodb       =>  /pro/beacon-2.0.0/BEACON/bin/bff2mongodb.py
Info: browser           =>  /pro/beacon-2.0.0/browser
```

...

```
Info: zip               =>  /usr/bin/pigz -p 1
Info:
Info: STARTING BEACON FUN
Info: **********************************
Info: Running the pipeline VCF2BFF...
Info: **********************************
Info: BEACON FINISHED OK
Info: Zay Gesunt
```

Data ingestion consists of 3 steps

3. Load data into MongoDB



B2RI: BFF → MongoDB

```
./beacon mongodb -p parameters.yaml
```

```
Info:
Info: Beacon 2.0.0
Info: Beacon exe: /pro/beacon-2.0.0/beacon
Info: (C) 2021 Manuel Rueda, PhD
Info: GNU General Public License v3
Info:
Info: ARGUMENTS USED:
Info: mode: mongodb
Info: --p parameters.yaml
Info:
Info: CONFIGURATION VALUES:
Info: bash4bff          =>  /pro/beacon-2.0.0/BEACON/bin/run_vcf2bff.sh
Info: bash4html         =>  /pro/beacon-2.0.0/BEACON/bin/run_bff2html.sh
Info: bash4mongodb      =>  /pro/beacon-2.0.0/BEACON/bin/run_bff2mongodb.sh
Info: bcftools          =>  /pro/NGSutils/bcftools-1.11/bcftools
Info: bff2json          =>  /pro/beacon-2.0.0/BEACON/bin/bff2json.pl
Info: bff2mongodb       =>  /pro/beacon-2.0.0/BEACON/bin/bff2mongodb.py
Info: browser           =>  /pro/beacon-2.0.0/browser
Info: hg19cosmic        =>  /media/mrueda/4TB/Databases/snpeff/v5.0/hg19/CosmicCodingMuts.normal.hg19.vcf.gz
Info: hg19dbnsfp        =>  /media/mrueda/4TB/Databases/snpeff/v5.0/hg19/dbNSFP4.1a_hg19.txt.gz
Info: hg19fasta         =>  /media/mrueda/4TB/Databases/genomes/ucsc.hg19.fasta.gz
Info: hg38cosmic        =>  /media/mrueda/4TB/Databases/snpeff/v5.0/hg38/CosmicCodingMuts.normal.hg38.vcf.gz
```

...

```
Info: STARTING BEACON FUN
Info: ********************************
Info: Running the pipeline BFF2MONGODB...
Info: ********************************
Info: BEACON FINISHED OK
```

Summary

- Beacon v2 is a specification
- B2RI is a Linux-based software developed by CRG
  - Data ingestion → NonSQL ←→ API
  - CINECA synthetic cohort EUROPE UK1
- BFF is a set of 7 files (JSON arrays) that match Beacon v2 Models

https://github.com/mrueda/Beacon2/tree/main/CINECA_synthetic_cohort_EUROPE_UK1

Beacon Instance and response format

- Beacon Instance (GiaB)
  - `https://ga4gh-approval-beacon-test.ega-archive.org/api`
- Beacon response (JSON) has 5 parts:
  - `meta, responseSummary, **response→ resultSets.results,** beaconHandovers`

GiAB instance is set as: **PUBLIC** and level **Record**

Query 1

GET all documents in endpoint <individuals>

https://ga4gh-approval-beacon-test.ega-archive.org/api/individuals


```
{"id":"","setType":"","exists":true,"resultsCount":3,"results":[{"_id":{"$oid":"61b21d2c0b2f27943afa85e7"}
,"diseases":[{"diseaseCode":{"id":"HP:000054","label":"intolerance"}}],"ethnicity":{"id":"HANCESTRO:0021",
"label":"Han Chinese"},"geographicOrigin":{"id":"GAZ:00002459","label":"United States of
America"},"id":"NA24631","phenotypicFeatures":[{"featureType":{"id":"NCIT:C3001","label":"NCIT:C3001"}}],"
sex":{"id":"NCIT:C46112","label":"male"}},

{"_id":{"$oid":"61b21d2c0b2f27943afa85e8"},"diseases":[{"diseaseCode":{"id":"HP:0008711","label":"prostati
c hyperplasia"}}],"ethnicity":{"id":"HANCESTRO:0021","label":"Han
Chinese"},"geographicOrigin":{"id":"GAZ:00002459","label":"United States of
America"},"id":"NA24694","phenotypicFeatures":[{"featureType":{"id":"NCIT:C37967","label":"NCIT:C37967"}}]
,"sex":{"id":"NCIT:C46112","label":"male"}},

{"_id":{"$oid":"61b21d2c0b2f27943afa85e9"},"diseases":[{"diseaseCode":{"id":"DOID:6364","label":"migraine"
}}],"ethnicity":{"id":"HANCESTRO:0021","label":"Han
Chinese"},"geographicOrigin":{"id":"GAZ:00002459","label":"United States of
America"},"id":"NA24695","phenotypicFeatures":[{"featureType":{"id":"HP:0000938","label":"HP:0000938"}}],"
sex":{"id":"NCIT:C16576","label":"female"}}
```


3 documents