BIPMed: Bioinformatics challenges and solutions

Benilton S Carvalho, Ph.D.

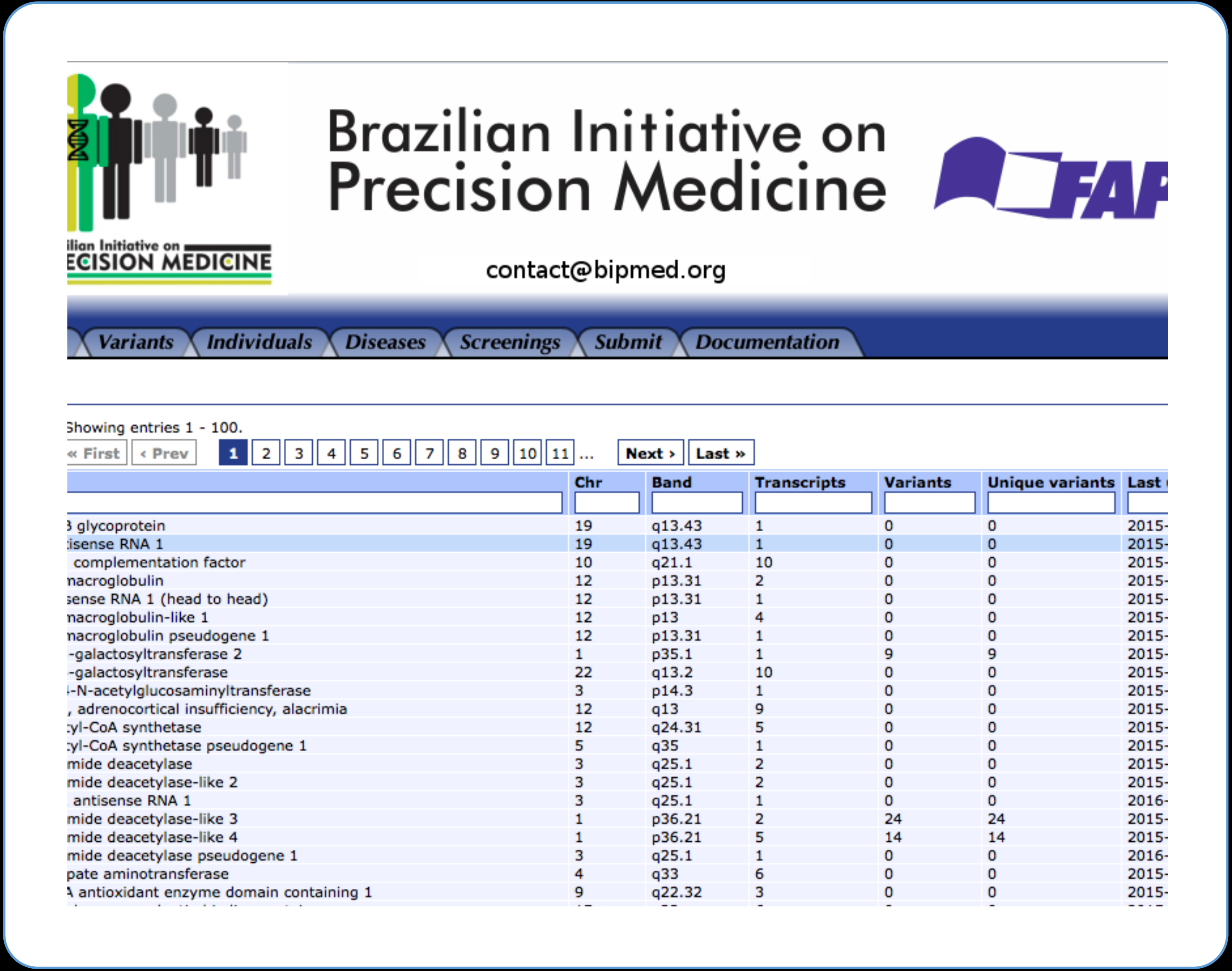
Institute of Mathematics, Statistics and Scientific Computing

University of Campinas

BIPMed Brazilian Initiative on PRECISION MEDICINE

Genomic Databases

- LOVD interface offers a lot of extra annotation;
- Loading data to this system requires a lot of computational resources;
- DB performance may be sub-optimum on the client-side;
- Testing transition to column-based DBMS;



contact@bipmed.org

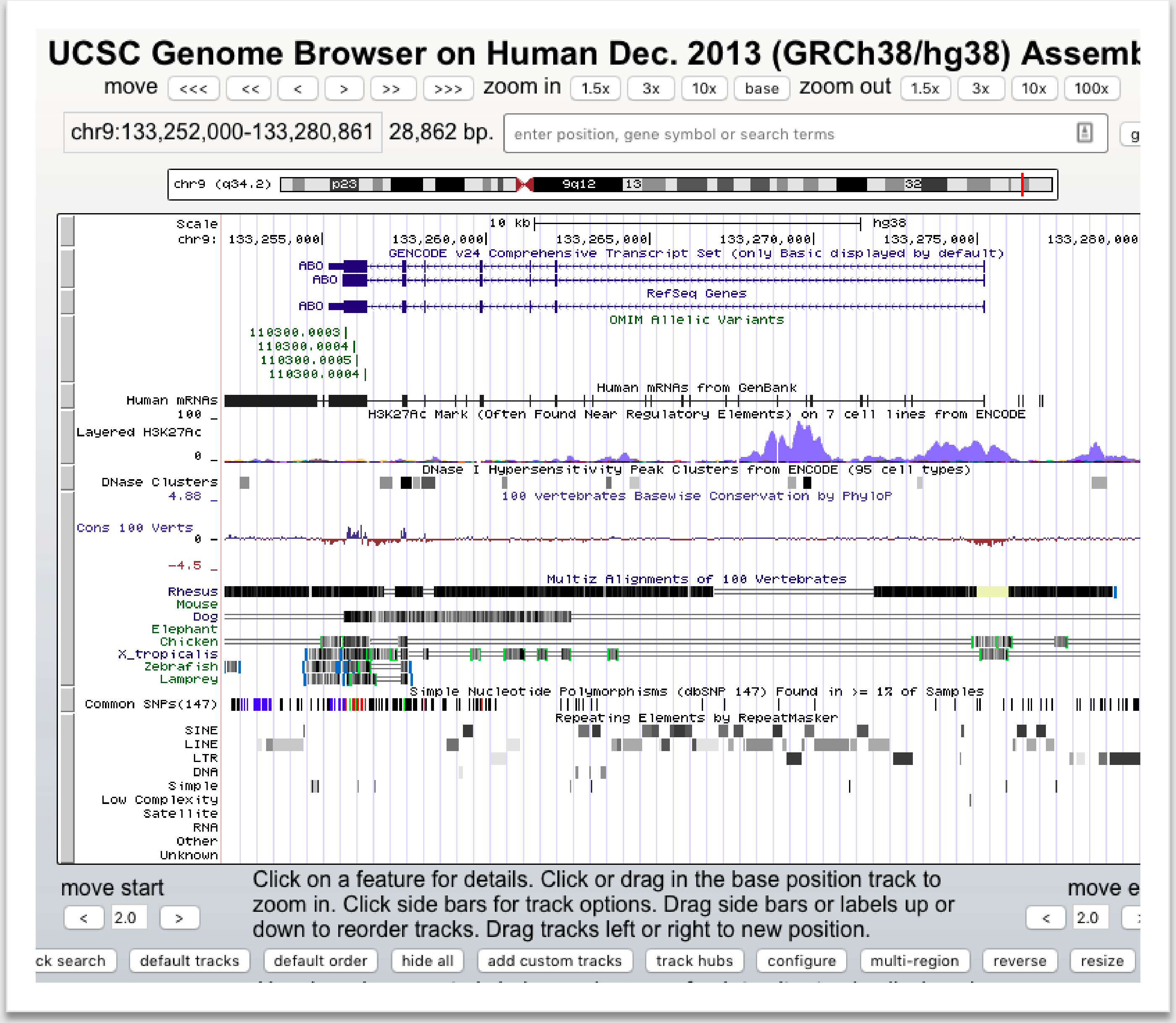
Showing entries 1 - 100.

	Chr	Band	Transcripts	Variants	Unique variants	Last
3 glycoprotein	19	q13.43	1	0	0	2015-
isense RNA 1	19	q13.43	1	0	0	2015-
complementation factor	10	q21.1	10	0	0	2015-
nacroglobulin	12	p13.31	2	0	0	2015-
sense RNA 1 (head to head)	12	p13.31	1	0	0	2015-
nacroglobulin-like 1	12	p13	4	0	0	2015-
nacroglobulin pseudogene 1	12	p13.31	1	0	0	2015-
-galactosyltransferase 2	1	p35.1	1	9	9	2015-
-galactosyltransferase	22	q13.2	10	0	0	2015-
-N-acetylglucosaminyltransferase	3	p14.3	1	0	0	2015-
, adrenocortical insufficiency, alacrimia	12	q13	9	0	0	2015-
yl-CoA synthetase	12	q24.31	5	0	0	2015-
yl-CoA synthetase pseudogene 1	5	q35	1	0	0	2015-
mide deacetylase	3	q25.1	2	0	0	2015-
mide deacetylase-like 2	3	q25.1	2	0	0	2015-
antisense RNA 1	3	q25.1	1	0	0	2016-
mide deacetylase-like 3	1	p36.21	2	24	24	2015-
mide deacetylase-like 4	1	p36.21	5	14	14	2015-
mide deacetylase pseudogene 1	3	q25.1	1	0	0	2016-
pate aminotransferase	4	q33	6	0	0	2015-
A antioxidant enzyme domain containing 1	9	q22.32	3	0	0	2015-

BIPMed Brazilian Initiative on PRECISION MEDICINE

Integration with existing tools

- Creating custom-tracks for use with UCSC Genome Browser;
  - All variants;
  - All variants found in at least 1%;
  - All variants found in at least 5%;
- Setting up a local infra-structure to accommodate BIPMed data via UCSC Genome Browser;



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assemb

chr9:133,252,000-133,280,861 28,862 bp.

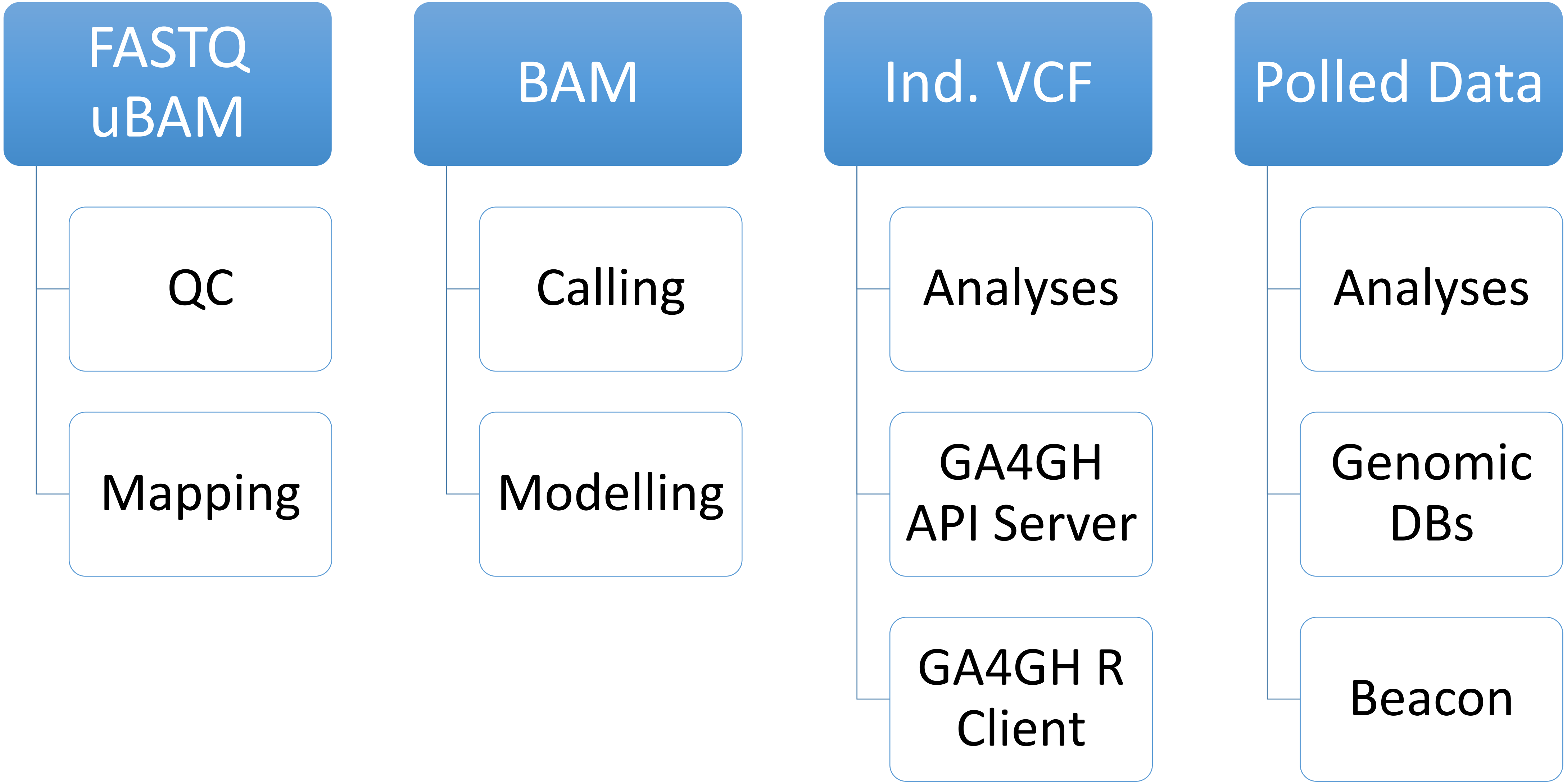
enter position, gene symbol or search terms

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

BIPMed Brazilian Initiative on PRECISION MEDICINE

BIPMed Brazilian Initiative on PRECISION MEDICINE

Data Management

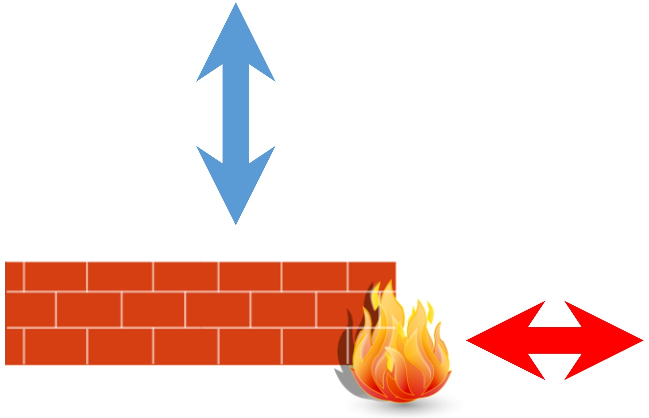


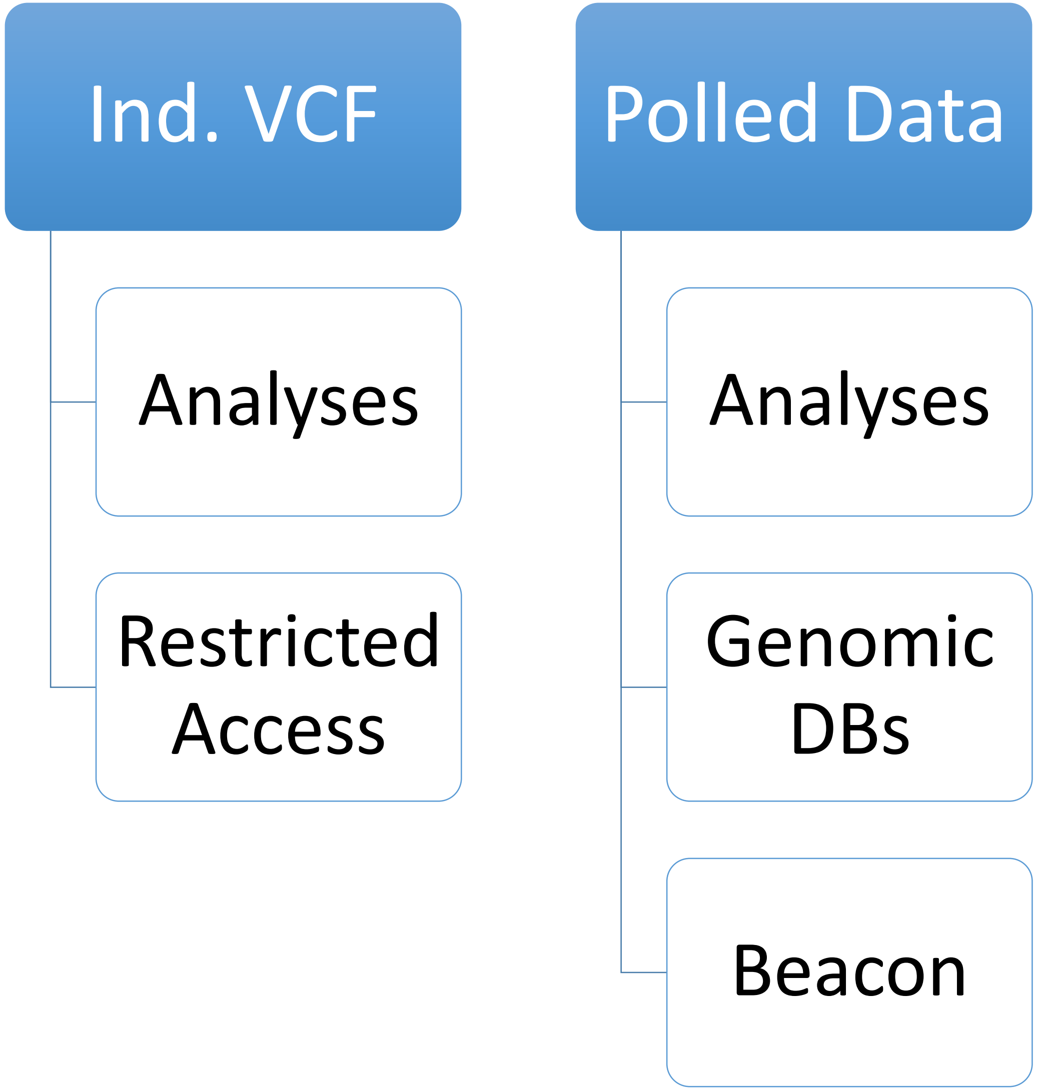
- FASTQ uBAM
  - QC
  - Mapping
- BAM
  - Calling
  - Modelling
- Ind. VCF
  - Analyses
  - GA4GH API Server
  - GA4GH R Client
- Polled Data
  - Analyses
  - Genomic DBs
  - Beacon

Data Management – Security – API

```
int iLength, iN;
double dblTemp;
bool again = true;

while (again) {
    iN = -1;
    again = false;
    getline(cin, sInput);
    system("cls");
    stringstream(sInput) >> dblTemp;
    iLength = sInput.length();
    if (iLength < 4) {
        again = true;
        continue;
    } else if (sInput[iLength - 3] != '.') {
        again = true;
        continue;
    } while (++iN < iLength) {
        if (isdigit(sInput[iN])) {
            continue;
            (iN == (iLength - 3) ) {
```





Ind. VCF

Analyses

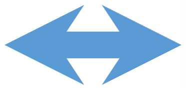
Restricted Access

Polled Data

Analyses

Genomic DBs

Beacon



```
double dblTemp;
bool again = true;

while (again) {
    iN = -1;
    again = false;
    getline(cin, sInput);
    system("cls");
    stringstream(sInput) >> dblTemp;
    iLength = sInput.length();
    if (iLength < 4) {
        again = true;
        continue;
    } else if (sInput[iLength - 3] != '.') {
        again = true;
        continue;
    } while (++iN < iLength) {
        if (isdigit(sInput[iN])) {
            continue;
            (iN == (iLength - 3) ) {
```

BIPMed Brazilian Initiative on PRECISION MEDICINE

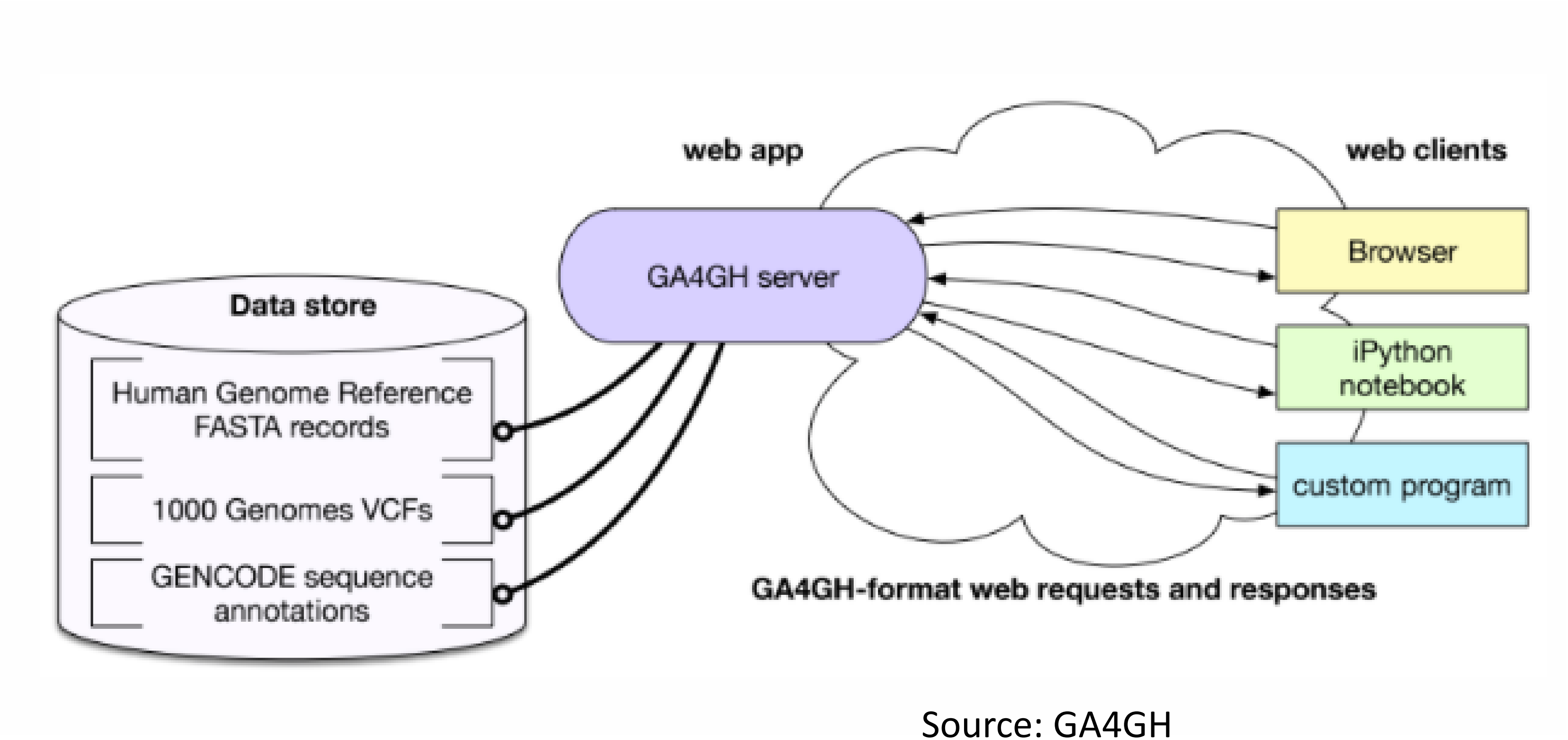
Programmatic Access – Databases

- DB infra-structure uses GA4GH Server API;
- We developed the GA4GH R Client to assist on programmatic access;
- DB access via GA4GH API is possible but not allowed as of today, as authentication is still being developed;
- Tools can be used internally;

GA4GH Server API

- The ga4gh-server software implements data schema for representing genomic variation data.
- It provides an application program interface (API) through the internet.
- Example: http://1kgenomes.ga4gh.org/
- Project page: https://github.com/ga4gh/server

GA4GH Server API



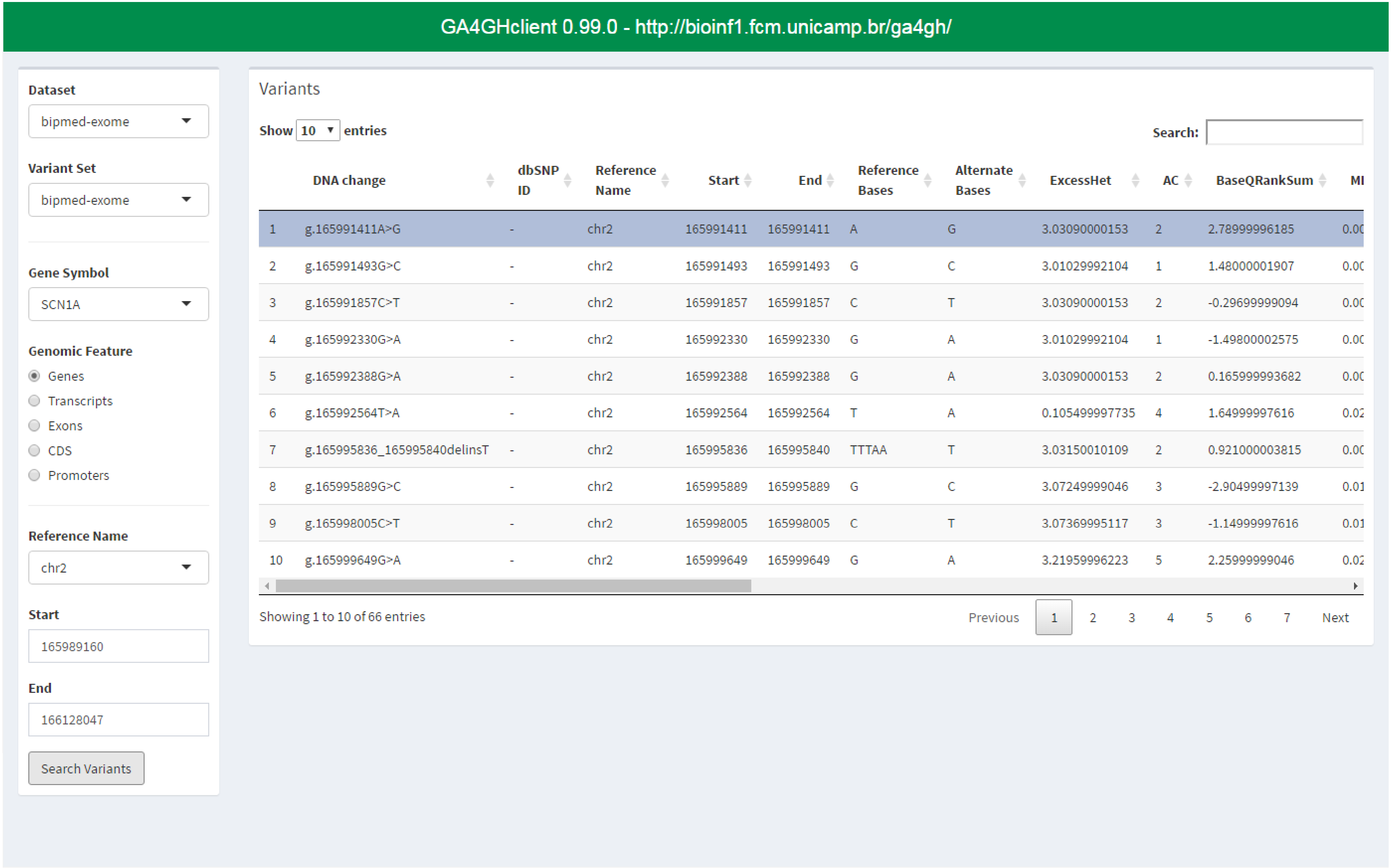
web app
web clients
Data store
Human Genome Reference FASTA records
1000 Genomes VCFs
GENCODE sequence annotations
GA4GH server
Browser
iPython notebook
custom program
GA4GH-format web requests and responses

Source: GA4GH

GA4GHclient package

- GA4GHclient is a Bioconductor package for accessing GA4GH API data servers.
- Provides tools to perform programmatic access to genomic data servers that follow the GA4GH standards.
- Integrates with other Bioconductor packages allowing creation of complex genomic data analysis.
- Also provides a web application to interact with genomic data.
- Project page: https://github.com/labbcb/GA4GHclient

GA4GH R Client



GA4GHclient 0.99.0 - http://bioinf1.fcm.unicamp.br/ga4gh/

Variants

Show 10 entries

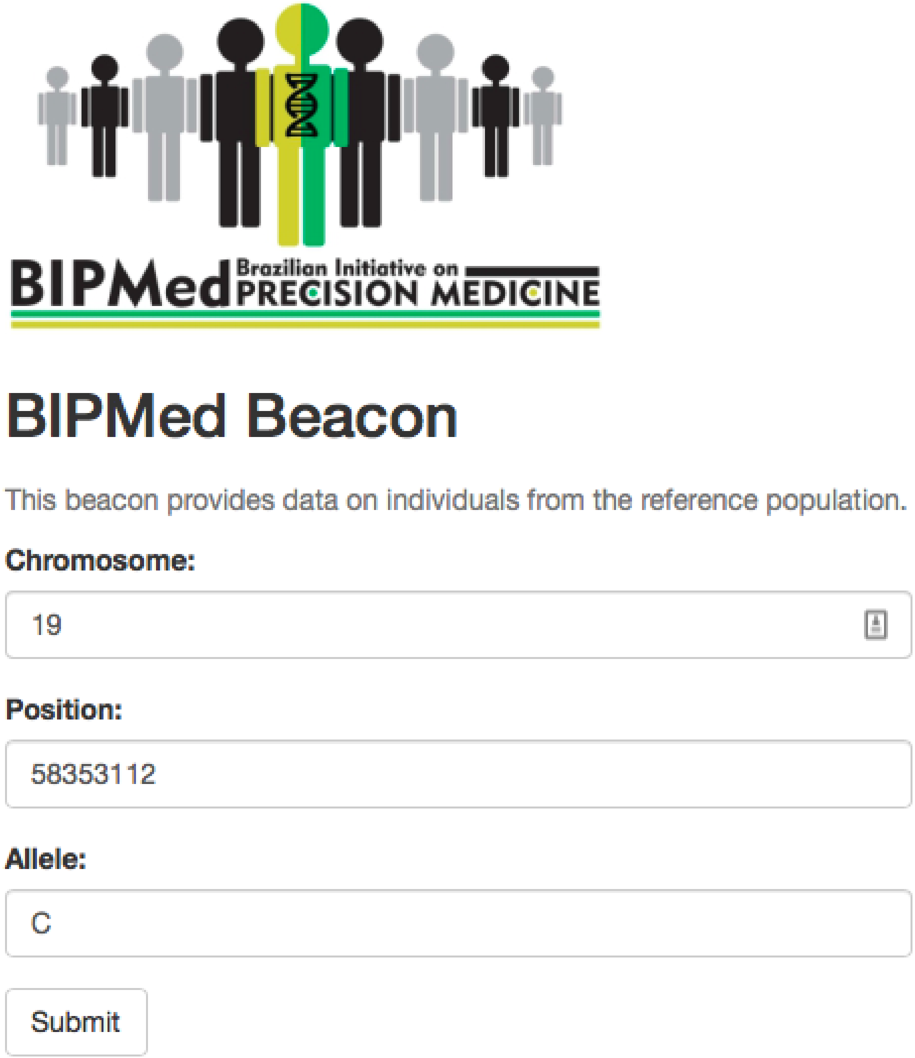
Search:

	DNA change	dbSNP ID	Reference Name	Start	End	Reference Bases	Alternate Bases	ExcessHet	AC	BaseQRankSum	MI
1	g.165991411A>G	-	chr2	165991411	165991411	A	G	3.03090000153	2	2.78999996185	0.0
2	g.165991493G>C	-	chr2	165991493	165991493	G	C	3.01029992104	1	1.48000001907	0.0
3	g.165991857C>T	-	chr2	165991857	165991857	C	T	3.03090000153	2	-0.29699999094	0.0
4	g.165992330G>A	-	chr2	165992330	165992330	G	A	3.01029992104	1	-1.49800002575	0.0
5	g.165992388G>A	-	chr2	165992388	165992388	G	A	3.03090000153	2	0.165999993682	0.0
6	g.165992564T>A	-	chr2	165992564	165992564	T	A	0.105499997735	4	1.64999997616	0.02
7	g.165995836_165995840delinsT	-	chr2	165995836	165995840	TTTAA	T	3.03150010109	2	0.921000003815	0.0
8	g.165995889G>C	-	chr2	165995889	165995889	G	C	3.07249999046	3	-2.90499997139	0.01
9	g.165998005C>T	-	chr2	165998005	165998005	C	T	3.07369995117	3	-1.14999997616	0.01
10	g.165999649G>A	-	chr2	165999649	165999649	G	A	3.21959996223	5	2.25999999046	0.02

Showing 1 to 10 of 66 entries

BIPMed Brazilian Initiative on PRECISION MEDICINE

BIPMed Beacon and Beacon Network



BIPMed Brazilian Initiative on PRECISION MEDICINE

BIPMed Beacon

This beacon provides data on individuals from the reference population.

Chromosome:

19

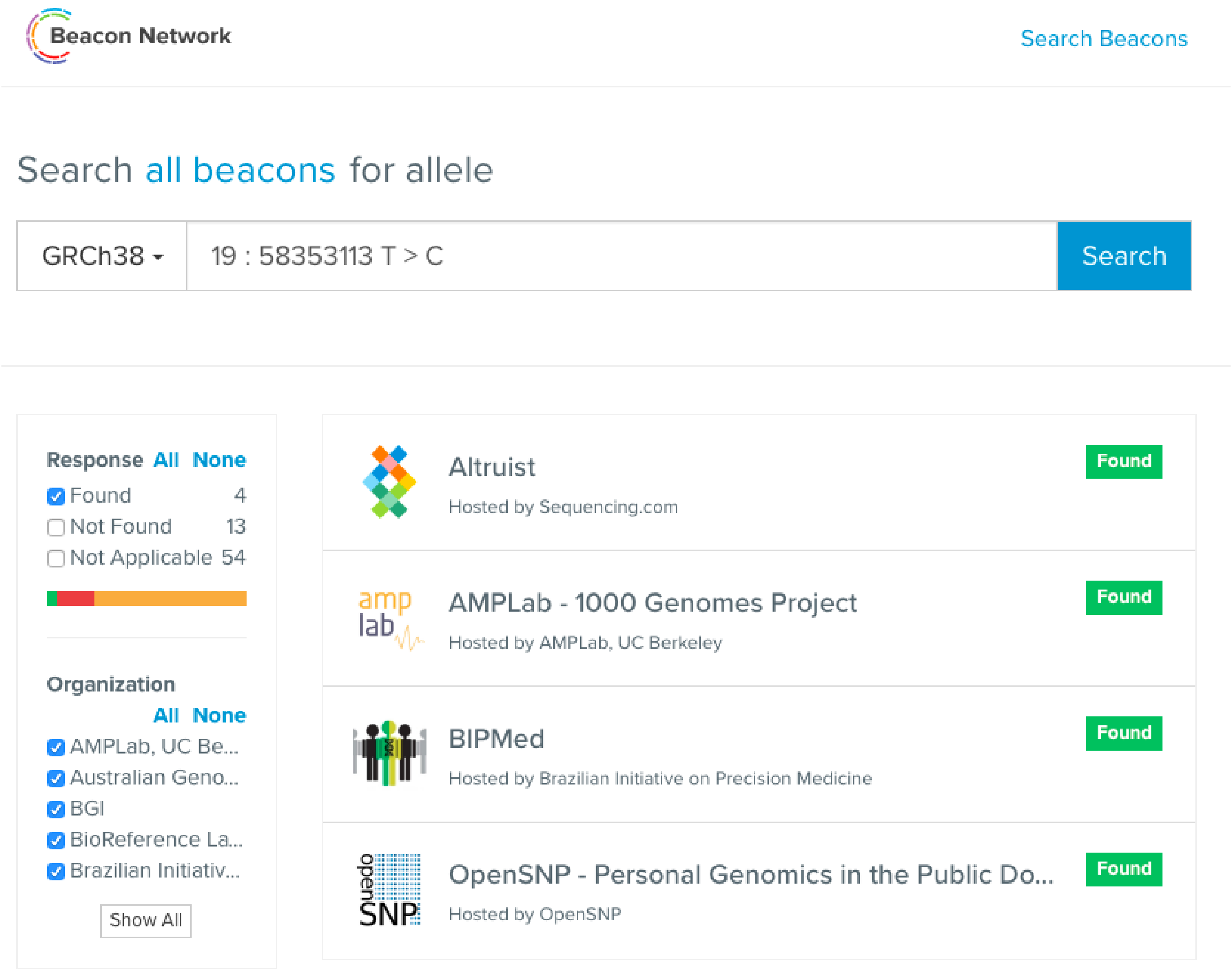
Position:

58353112

Allele:

C

Submit



Beacon Network

Search Beacons

Search all beacons for allele

GRCh38 ▾ | 19 : 58353113 T > C | Search

Response All None

☑ Found 4

☐ Not Found 13

☐ Not Applicable 54

Organization

All None

☑ AMPLab, UC Be...

☑ Australian Geno...

☑ BGI

☑ BioReference La...

☑ Brazilian Initiativ...

Show All

Altruist
Hosted by Sequencing.com
Found

AMPLab - 1000 Genomes Project
Hosted by AMPLab, UC Berkeley
Found

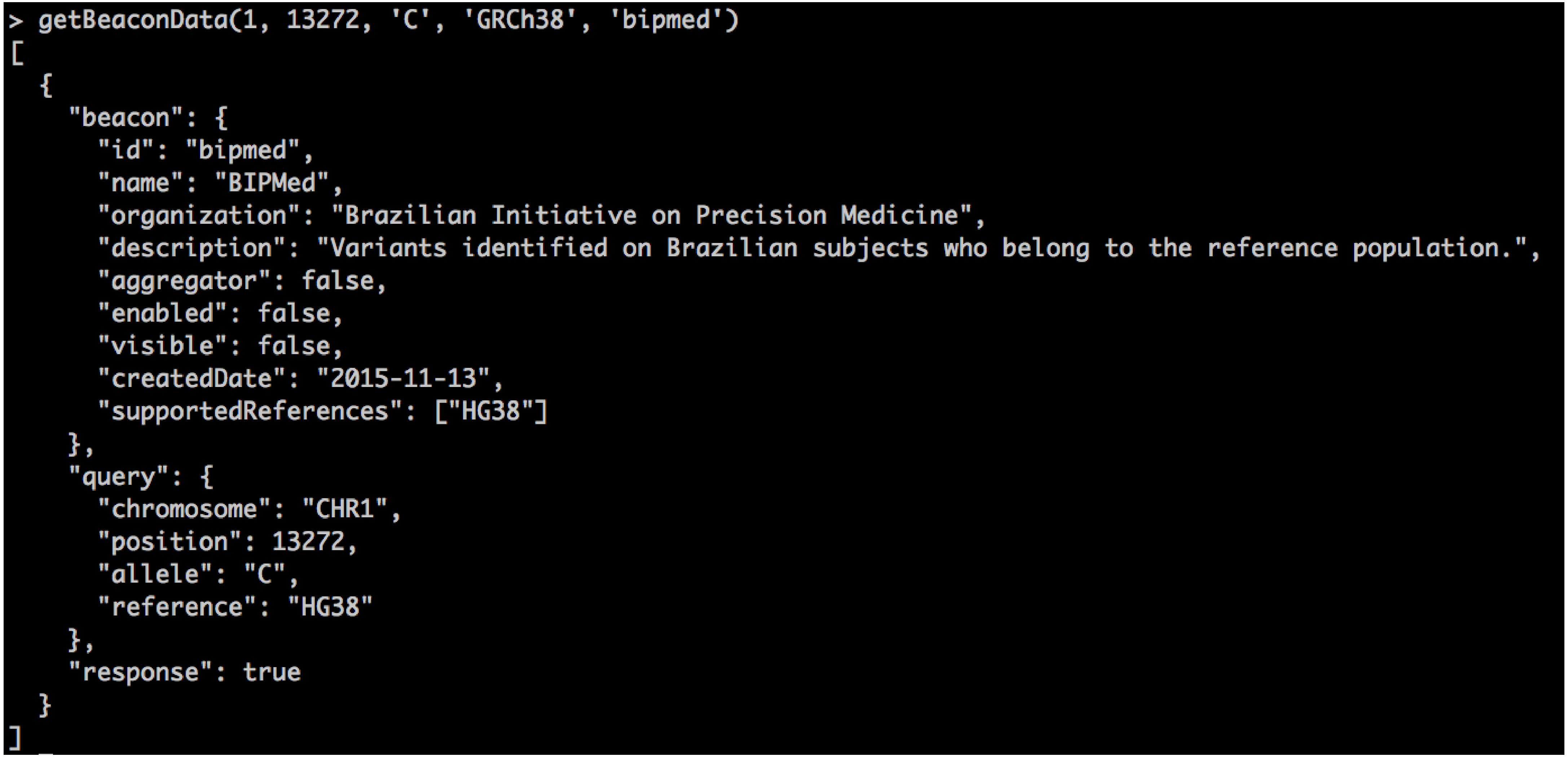
BIPMed
Hosted by Brazilian Initiative on Precision Medicine
Found

OpenSNP - Personal Genomics in the Public Do...
Hosted by OpenSNP
Found

Programmatic Access – Beacon

- BIPMed Beacon connected to Beacon Network;
- Beacon Network can be accessed via Beacon API;
- Beacon API can access BIPMed Beacon directly;
- Tools are under development and designed to allow for queries through computer programs, assisting on variant filtering;

Programmatic Access – Beacon

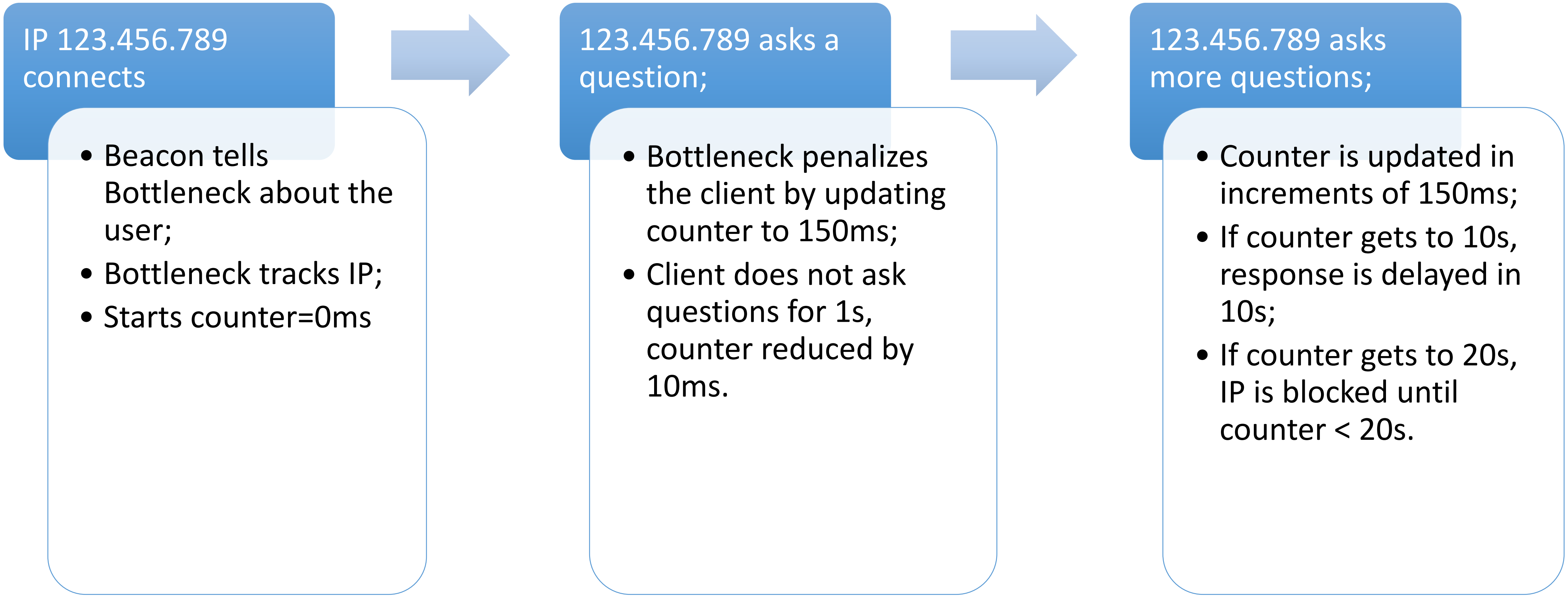


```
> getBeaconData(1, 13272, 'C', 'GRCh38', 'bipmed')
[
  {
    "beacon": {
      "id": "bipmed",
      "name": "BIPMed",
      "organization": "Brazilian Initiative on Precision Medicine",
      "description": "Variants identified on Brazilian subjects who belong to the reference population.",
      "aggregator": false,
      "enabled": false,
      "visible": false,
      "createdDate": "2015-11-13",
      "supportedReferences": ["HG38"]
    },
    "query": {
      "chromosome": "CHR1",
      "position": 13272,
      "allele": "C",
      "reference": "HG38"
    },
    "response": true
  }
]
```

Bottleneck Server

- To avoid genome-wide queries, the BIPMed Beacon uses IP Throttling;
- IP Throttling is performed by the Bottleneck Server, as used by UCSC;
- In practice, the Beacon slows down requests when too many come from the same IP.

Bottleneck Server



IP 123.456.789 connects

- Beacon tells Bottleneck about the user;
- Bottleneck tracks IP;
- Starts counter=0ms

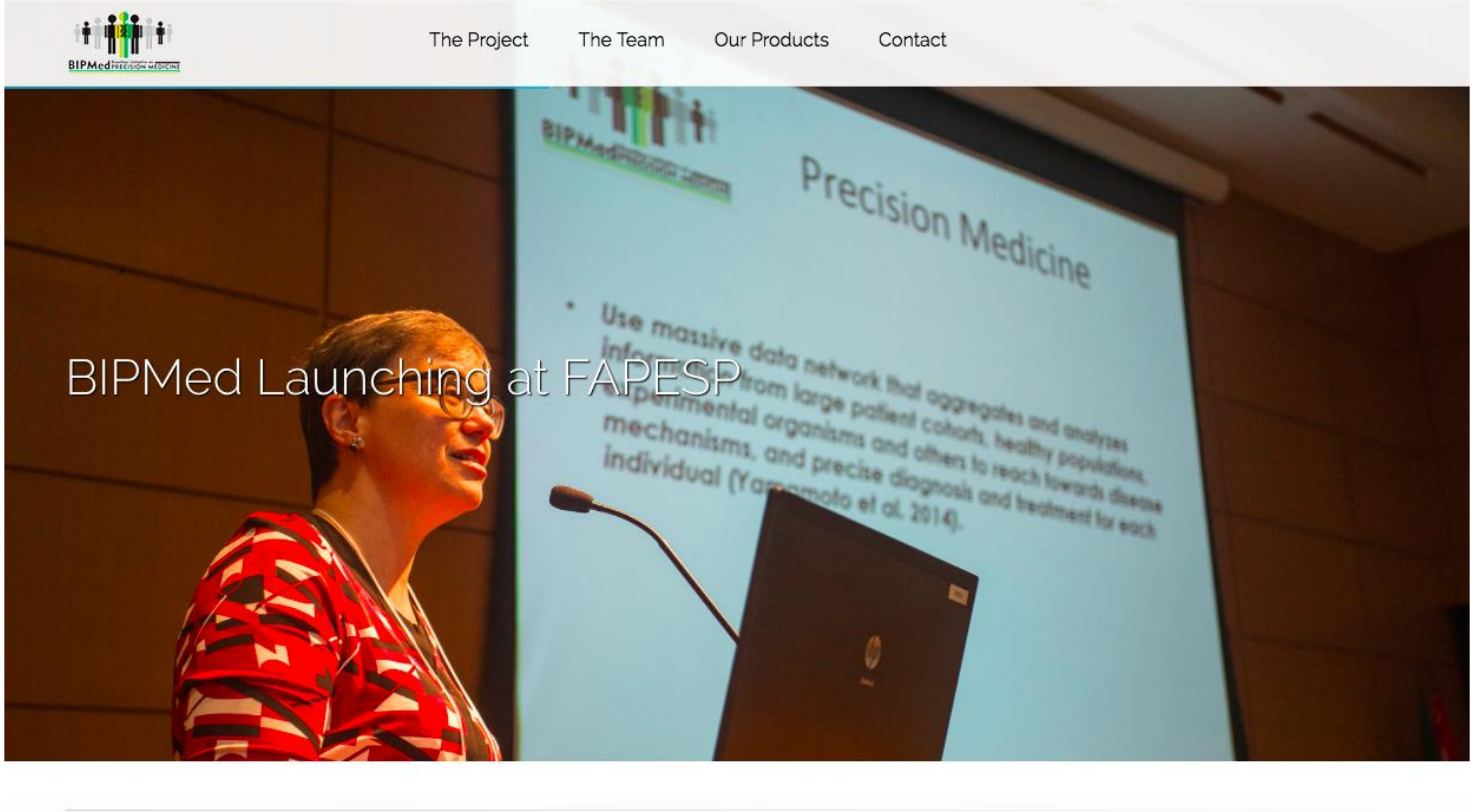
123.456.789 asks a question;

- Bottleneck penalizes the client by updating counter to 150ms;
- Client does not ask questions for 1s, counter reduced by 10ms.

123.456.789 asks more questions;

- Counter is updated in increments of 150ms;
- If counter gets to 10s, response is delayed in 10s;
- If counter gets to 20s, IP is blocked until counter < 20s.

BIPMed Website
http://www.bipmed.org


BIPMed Brazilian Initiative on PRECISION MEDICINE
The Project
The Team
Our Products
Contact
Precision Medicine
Use massive data network that aggregates and analyzes information from large patient cohorts, healthy populations, experimental organisms and others to reach towards disease mechanisms, and precise diagnosis and treatment for each individual (Yamamoto et al. 2014).

BIPMed Launching at FAPESP

Website security

- Website is secured by Cloudfare;
- Contents are copied across many Cloudfare servers around the world;
- Every request to the website is initially sent to the Cloudfare server;
- Cloudfare analyzes requests and decides whether or not they come from a client flagged as a threat;

Website security

- If client is not a threat:
  - Cloudfare finds the server that is the closest to the client;
  - This server replies to the request by sending cached versions of the website;
  - Connection is much faster and our server does not need to handle traffic;
- If client is a threat:
  - Cloudfare will not reply to requests.
  - Our “original” server is protected from attacks.
- All services provided by us may use a similar system.

Reliability

- Genomic databases and Beacons are mirrored;
- Current mirrors:
  - USP
  - School of Medicine – UNICAMP;
  - Institute of Chemistry – UNICAMP;
- Mirrors are designed to work as backup systems in case of problems with the main server.

BIPMed Brazilian Initiative on PRECISION MEDICINE

Thank you!

- Visit our website: http://www.bipmed.org
- Email us: contact@bipmed.org