ClinGen Allele Registry

Protein Variants

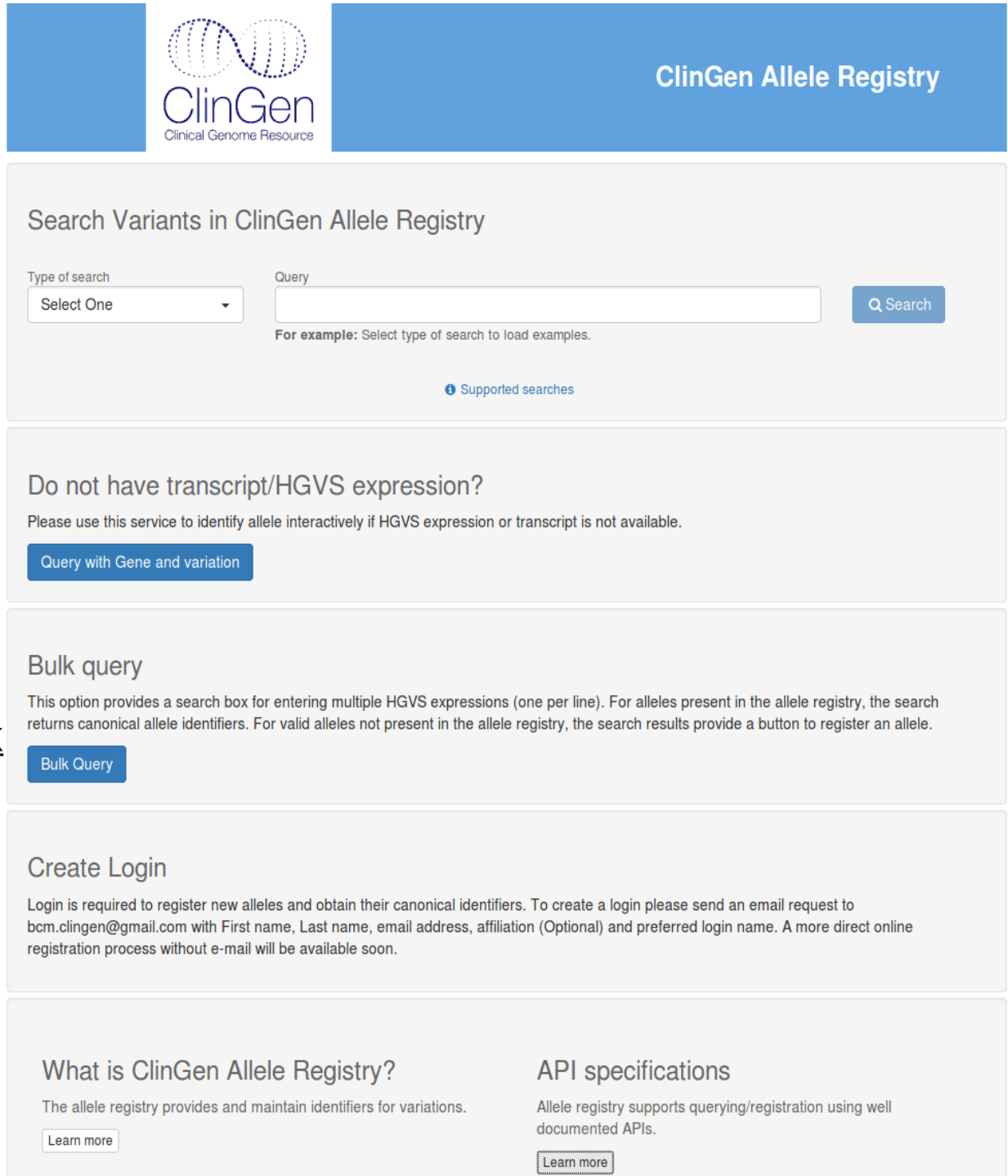
Piotr Pawliczek

April 2017

Current version

Last CAR release (0.10.04) is installed on the Rackspace servers.

- New User Interface (beta version)
- Production server: http://reg.genome.network http://reg.clinicalgenome.org
- Test server: http://reg.test.genome.network
- REST API with JSON-LD responses
- Documentation for API with examples in Python and Ruby



Next version - in development

Next CAR release (0.11.x) will appear in ~3-4 weeks on the test server (http://reg.test.genome.network).

- ~250 millions of Canonical Alleles (vs. ~15 millions in 0.10.x)
- Identifiers and links to dbSNP, gnomAD, myVariant.info
- Larger & faster bulk queries & registrations
- API – new parameters to control format of output documents
- Better support for protein variants (to discuss)

Genomic alleles

Single returned page in UI (or document in API) corresponds to one Canonical Allele. It contains list of corresponding Contextual Alleles and known external identifiers (and links).

ClinGen Clinical Genome Resource

ALLELE REGISTRY

CA000123

- Type: Nucleotide
- dbSNP: 11540654
- ClinVar Alleles: NM_000546.5(TP53):c.329G>A (p.Arg110His), allele ID = 133265
- ClinVar Variations: 127808, RCV = RCV000122182, RCV000115719
- ExAC: 17:7579358 C / T

Genomic Alleles

HGVS	Reference Sequence	Position	Allele
NG_017013.2:g.16511G>A , LRG_321:g.16511G>A		16510-16511	A
NC_000017.9:g.7520083C>T	NCBI36, chr 17	7520082-7520083	T
NC_000017.11:g.7676040C>T , CM000679.2:g.7676040C>T	GRCh38, chr 17	7676039-7676040	T
NC_000017.10:g.7579358C>T , CM000679.1:g.7579358C>T	GRCh37, chr 17	7579357-7579358	T

Transcript Alleles

HGVS	Reference Sequence	Position	Allele
NM_001276761.1:c.212G>A		527-528	A
NM_001276760.1:c.212G>A		530-531	A
NM_001276696.1:c.212G>A		530-531	A
NM_001276695.1:c.212G>A		530-531	A
NM_001126118.1:c.212G>A , LRG_321t8:c.212G>A		647-648	A
NM_001126114.2:c.329G>A , LRG_321t3:c.329G>A		530-531	A
NM_001126113.2:c.329G>A , LRG_321t4:c.329G>A		530-531	A
NM_001126112.2:c.329G>A , LRG_321t2:c.329G>A		527-528	A
NM_000546.5:c.329G>A , LRG_321t1:c.329G>A		530-531	A
ENST00000635293.1:c.212G>A		471-472	A
ENST00000622645.4:c.212G>A		461-462	A
ENST00000620739.4:c.212G>A		518-519	A
ENST00000619485.4:c.212G>A		464-465	A
ENST00000617185.4:c.329G>A		530-531	A

Protein alleles

Canonical Alleles with protein variants contain always one protein mutation (Contextual Allele). There is no link between genomic & protein alleles.

CA162562

See this allele in the new user interface (beta)

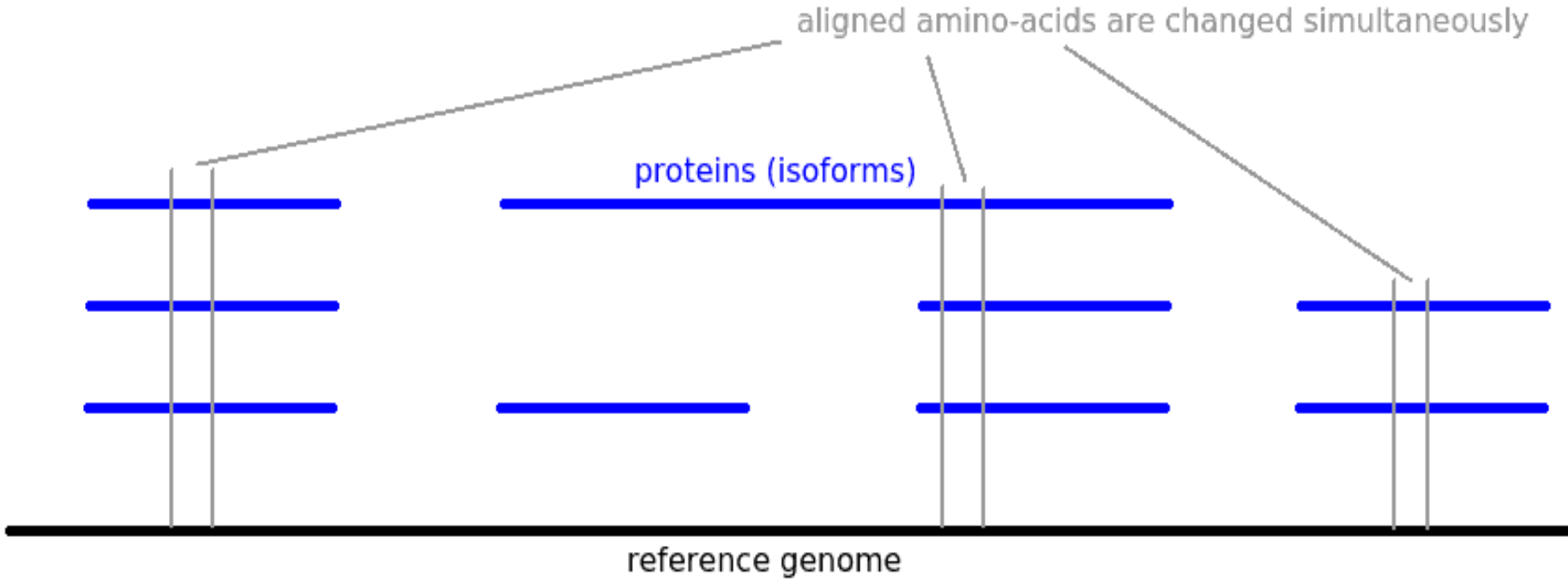
- Type: Amino-acid

Amino-Acid Alleles

HGVS	Reference Sequence	Position	Allele
NP_001119590.1:p.Arg71His		70-71	H

Grouping together protein alleles

- For each protein's amino-acid we can find corresponding codon on the reference genome
- For this codon we can find all proteins, that contain amino-acid produced from this codon (no frameshift is allowed)



Grouping together protein alleles

- Using this approach we can calculate alignment between proteins (isoforms).
- For given protein variant, a list of corresponding protein variants can be calculated.
- For given protein variant CAR can return page (document) with corresponding variants on different proteins, that always co-occur together (Canonical Protein Allele).
- Two Contextual Protein Allele belong to the same Canonical Protein Allele => they occur simultaneously (it does not work in the opposite direction).

Protein variant (modification) definition

- Protein variant (modification) definition:
  - Sequence modification (minimal, right aligned)
    - Range of replaced amino-acids (may be empty)
    - New amino-acids (may be empty)
  - Frameshift
    - 0 – no frameshift
    - 1 – skip 1bp
    - 2 – skip 2bp
- Complex Alleles can be aligned between proteins
- Several different protein variants with frameshifts can be translated to the same HGVS

Links between genomic & protein alleles

- Contextual Genomic Allele may have link to single Contextual Protein Allele (exception: two “active” start codons)
- Canonical Genomic Allele may have link to single Canonical Protein Allele
- Contextual Protein Allele may have links to several Contextual Genomic Allele
- Canonical Protein Allele may have links to several Canonical Genomic Allele

Others

- Change prefix of identifiers of Protein Alleles
- What should have unique ID?
  - Canonical Protein Allele
  - Contextual Protein Allele (this is probably needed)
  - both
- Where to bind external identifiers?
  - Canonical Protein Allele
  - Contextual Protein Allele
  - Both
- Can Contextual Protein Allele belong to two Canonical Protein Alleles?