The Transcript Archive

September 12, 2017

Matthew Laird
Ensembl, Core Team

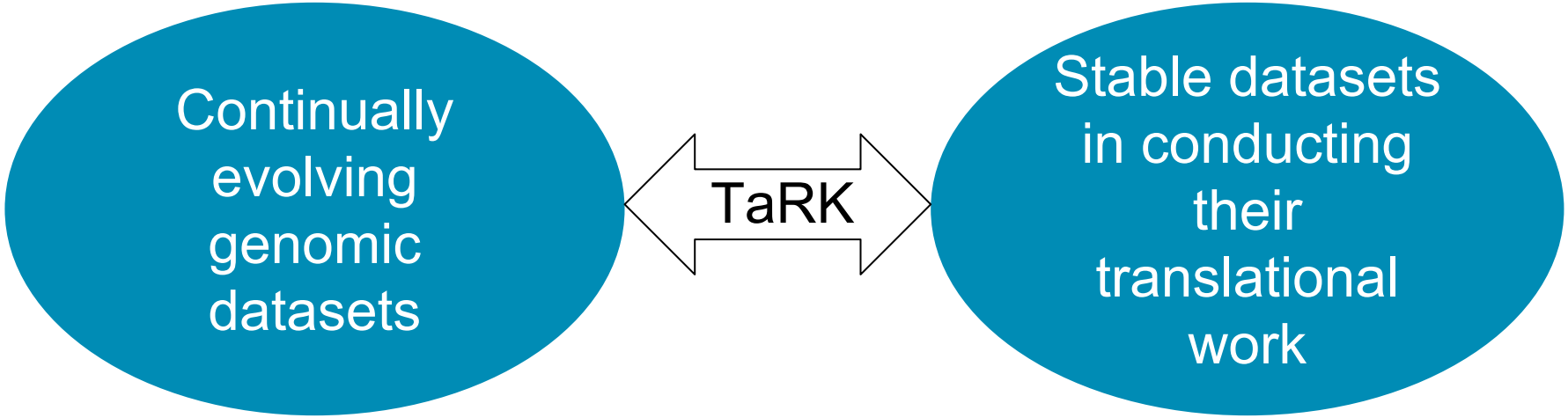
EMBL-EBI

Ensembl

Ensembl Updates - TaRK

Transforming Genomic Medicine Initiative

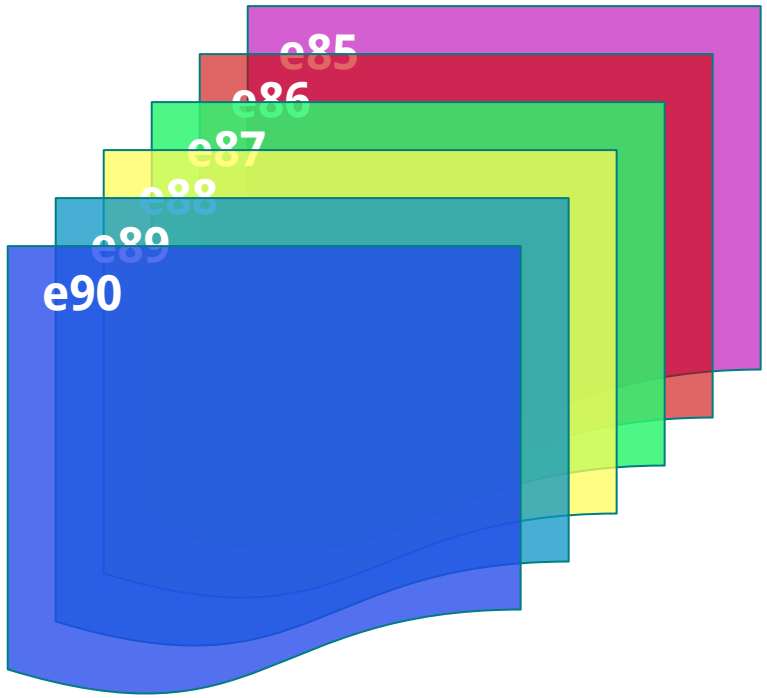
The challenge for researchers



- The need to tag a stable transcript set that won't change on researchers between releases

Ensembl Updates - Sequence Store

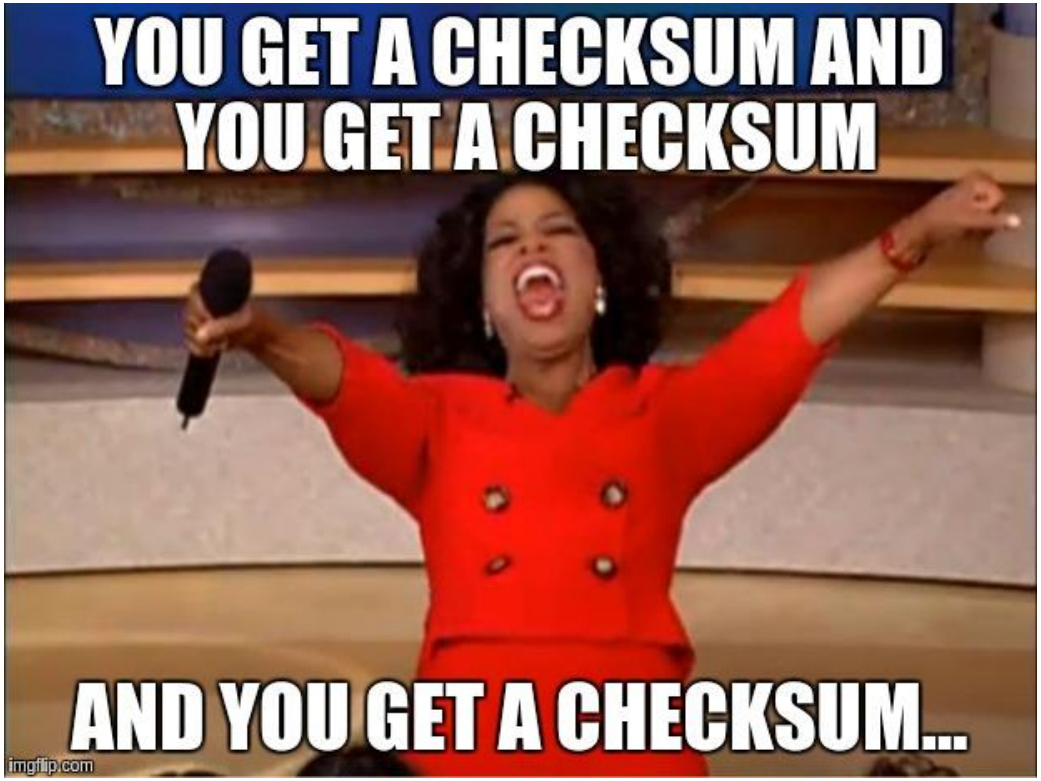
- With the growing number of genomes, a way to reduce redundancy between releases was needed
- Annotations could go 10 or more releases without updates to sequence or structure in Ensembl



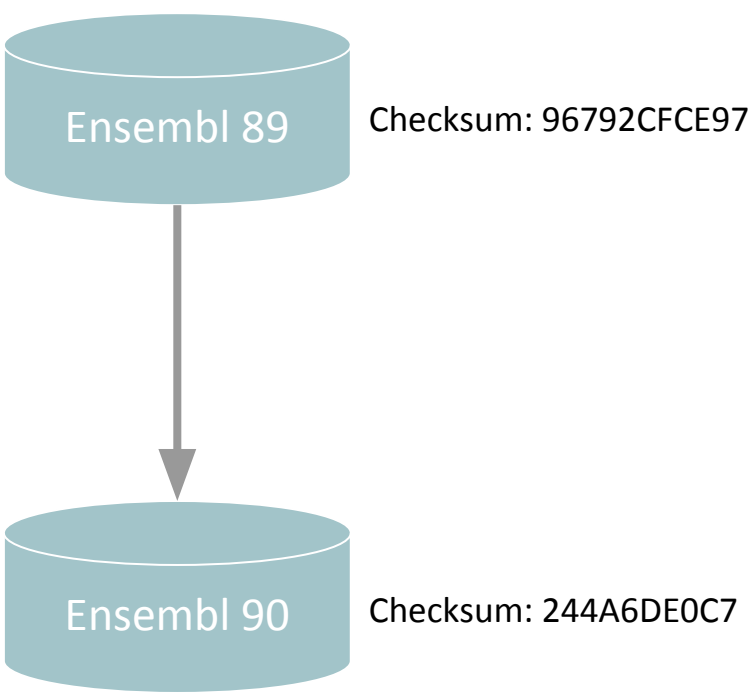
TaRK & Sequence Store Goals

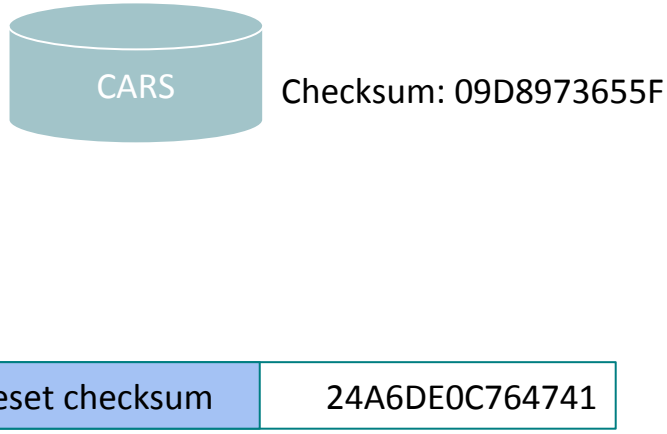
- Maintain persistent tagged transcript set across Ensembl releases
- Calculate differences between transcript set releases and Ensembl releases/data freezes
- Checksum datasets at multiple levels; from release level to individual feature
- RESTful interface, rich queries slicing over multiple tagsets, criteria in creating resultset

TaRK - Checksums



Set checksums

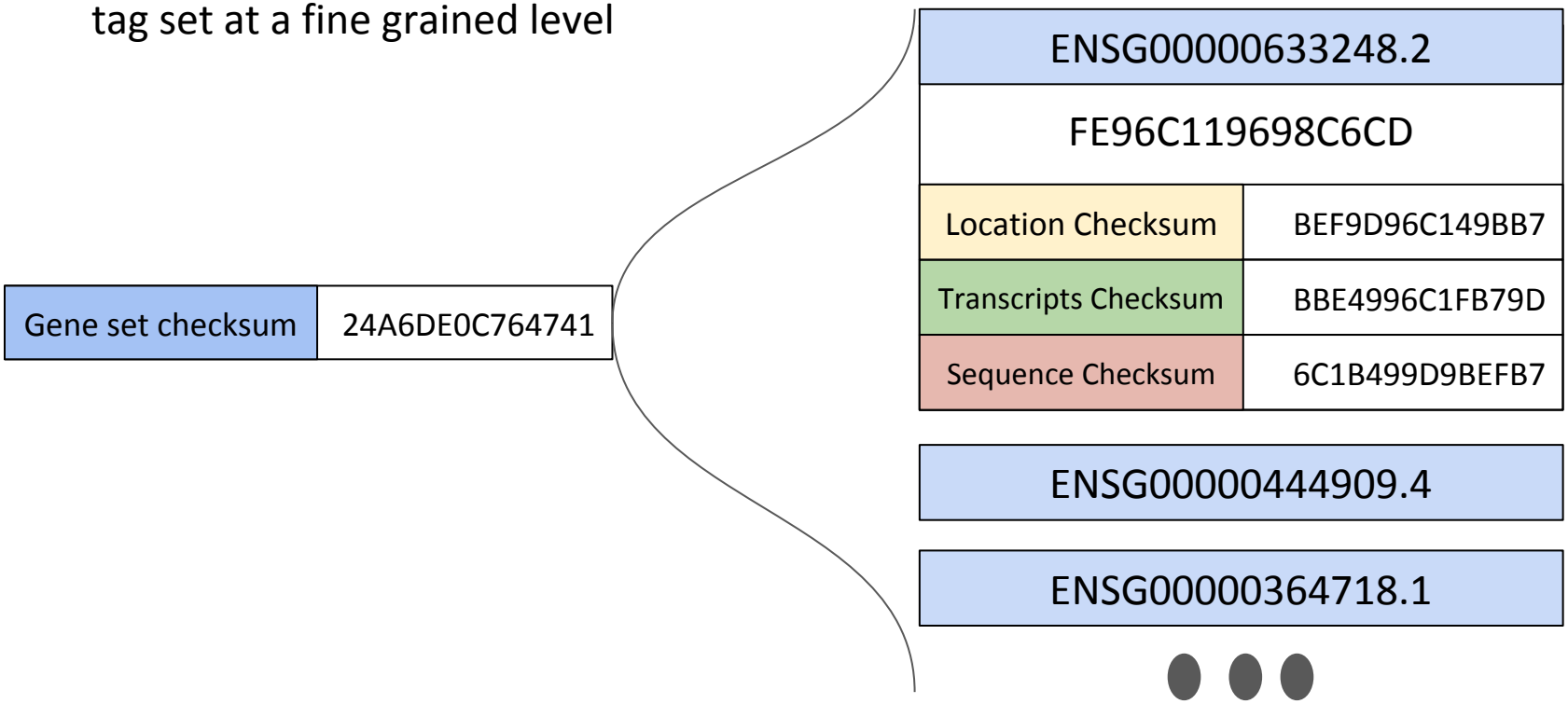




- Checksums currently SHA1 based, but we will be regenerating the database using SHA-512 in the near future

Set checksums

- Dig down through sets finding difference between release or tag set at a fine grained level



Revision tracking

- Dig down to find changes at sub-record level between two features

  eg. exon location shift

ENST00000444909.4	
7B6773A24743AA48	
Location Checksum	7BF3A24743AA48
Exon set Checksum	43AA487BF3A247
Sequence Checksum	E743AA4AF3AA48

ENST00000444909.5	
FE96C119698C6CD	
Location Checksum	7BF3A24743AA48
Exon set Checksum	6F0EB5E6A539
Sequence Checksum	E743AA4AF3AA48

Reduced redundancy

- Redundant sequences are only stored once
- This includes across assemblies and species

ENST00000444909.4	
7B6773A24743AA48	
Assembly	GRCh37
Location Checksum	7BF3A24743AA48
Exon set Checksum	43AA487BF3A247
Sequence Checksum	E743AA4AF3AA48

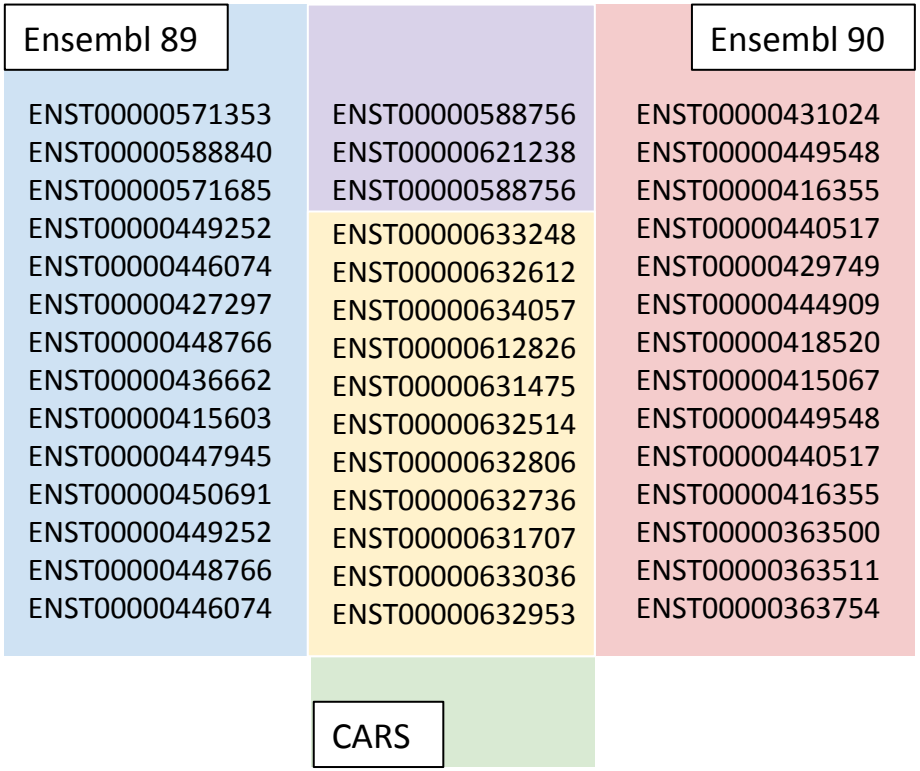
ENST00000444909.4	
7B6773A24743AA48	
Assembly	GRCh38
Location Checksum	7BF3A24743AA48
Exon set Checksum	43AA487BF3A247
Sequence Checksum	E743AA4AF3AA48

Sequence Checksum	E743AA4AF3AA48

CTGCAATCGACACACCCTAGCGGACAATTTTAACCC
TGTGTCTGAGGAGCGTGGCAAAGTTGCCAAGATTGT
TTCTACCTCTTTGAGATGGATAGCAGCCTGGCCTGT
TCACCAGAGATCTCAGCCACCTCAGTGTGGGTTCCA
TCTTACTTGTCCTGGTAGCAGATGGCTGTGACTTTG
TCTGTACCGTTCTAAACCTCGAAATGTGCCTGCAGC
ATATCGTGGTGTGGGGGATGACCAGCTGGGAC

Rich queries

- Rich queries taking advantage of multiple tagsets available
  - “All human transcripts identical between E89 and E90 and in CARS”



Ensembl Difference Set

Release comparison endpoint detailing changes between Ensembl releases

- Features added/removed
- Location and sequence changes
- Stable id remapping


```
{
  stable_id: "ENSG00000122877",
 ▾ version: {
      updated: 14,
      base: 13
  },
 ▾ transcript_differences: [
    ▾ {
      ▾ exon_differences: [
        ▾ {
            stable_id: "ENSE00001760892",
          ▾ version: {
              updated: 2,
              base: 1
            },
          ▾ location: {
              updated: "GRCh38:10:62816070:62816315:-1",
              base: "GRCh38:10:62816070:62816366:-1"
            },
          ▾ sequence: {
              updated: "GTTATAATAACACTACACCAGCAACTCCTGGCTTCCCAGCAGCCGGAACACAGACAGGAGAGAGTCAGTGGCAAATAGACATTTTTCTTATTTCTTAAAAAACAGCAACTTGTTTGCTACTTTTATTTCT
              base:
              "AACTGAGCGAGGAGCAATTGATTAATAGCTCGGCGAGGGGACTCACTGACTGTTATAATAACACTACACCAGCAACTCCTGGCTTCCCAGCAGCCGGAACACAGACAGGAGAGAGTCAGTGGCAAATAGACATTTTTCT
            }
        },
        ▾ {
            stable_id: "ENSE00000834013",
          ▾ missing: {
              release: "84"
            }
        },
        ▾ {
            stable_id: "ENSE00003792399",
          ▾ added: {
              release: "85"
            }
        }
      ],
      stable_id: "ENST00000411732",
    ▾ version: {
        updated: 3,
        base: 2
      },
    ▾ location: {
        updated: "GRCh38:10:62812000:62816315:-1",
        base: "GRCh38:10:62812003:62816366:-1"
      },
    ▾ sequence: {
        updated:
        "GTTATAATAACACTACACCAGCAACTCCTGGCTTCCCAGCAGCCGGAACACAGACAGGAGAGAGTCAGTGGCAAATAGACATTTTTCTTATTTCTTAAAAAACAGCAACTTGTTTGCTACTTTTATTTCTGTTGATTTTTTTTTCT
        base:
        "AACTGAGCGAGGAGCAATTGATTAATAGCTCGGCGAGGGGACTCACTGACTGTTATAATAACACTACACCAGCAACTCCTGGCTTCCCAGCAGCCGGAACACAGACAGGAGAGAGTCAGTGGCAAATAGACATTTTTCTTATTTCT
      }
    },
  ▾ {
      stable_id: "ENST00000637191",
    ▾ added: {
        release: "85"
      }
    }
  ]
},
```

Ensembl Difference Set

Intersect the two datasets to get a sense of what changes might occur in the frozen transcript set between Ensembl releases.

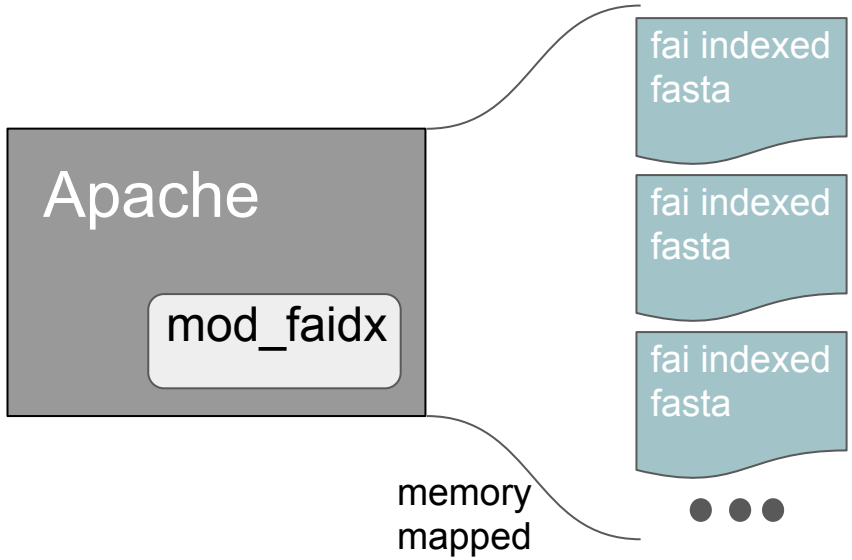
```
ENSG00000215474    SKOR2     protein_coding ENST00000620245
ENSG00000173621    LRFN4     protein_coding ENST00000309602
ENSG00000139624    CERS5     protein_coding ENST00000317551
ENSG00000157326    DHRS4     protein_coding ENST00000313250
ENSG00000087116    ADAMTS2   protein_coding ENST00000251582
```


```
{
  stable_id: "ENSG00000215474",
 ▾ version: {
      updated: 7,
      base: 6
  },
 ▾ location: {
      updated: "GRCh38:18:47206322:47251603:-1",
      base: "GRCh38:18:47212089:47249183:-1"
  },
 ▾ transcript_differences: [
    ▾ {
      ▾ sequence: {
            updated:
            "CGTACCCACACTTTCTGCGGTGGAGGGGACGCCCCGCCGCAATTCAGGCCGTCATTCTCCCCAGGCCGGGGTTTGAGCGCCATTCGCTCGGGCC
            base:
            "ATGGCTTCCAGTCCGCTGCCAGGGCCCAACGACATCCTGCTGGCGTCGCCGTCGAGCGCCTTCCAGCCCGACACGCTGAGCCAGCCGCGGCCAG
        },
      ▾ exon_differences: [
        ▾ {
            stable_id: "ENSE00003792160",
          ▾ added: {
              release: "85"
            }
          },
```

Reference sequence (mod_faidx)

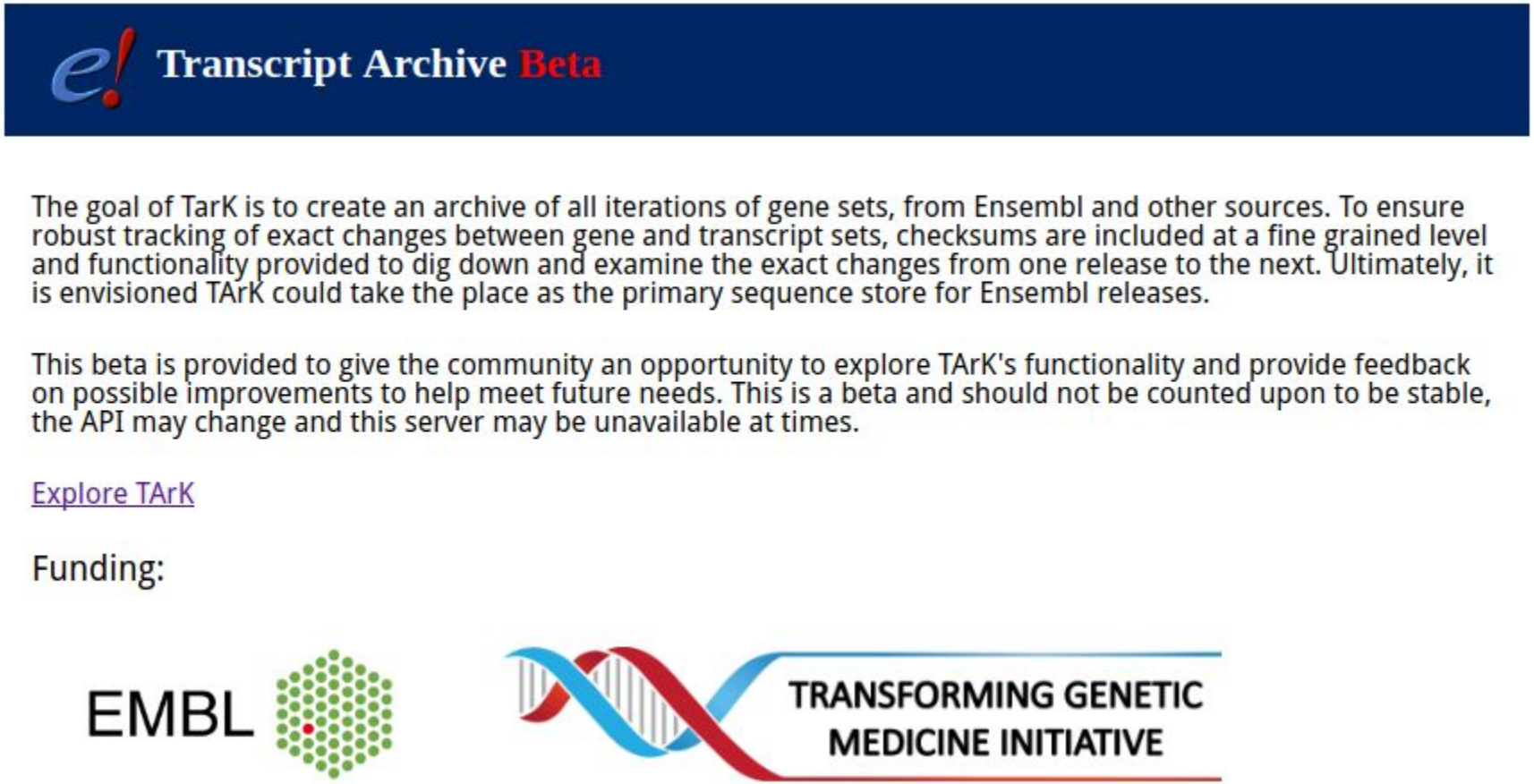
Bulk sequence storage and serving is another growth pinch point.

- Apache module written in C
- Uses htslib to memory map compressed, indexed fasta files
- Can return individual sequences, or assemble a series of coordinates and translate to protein sequence
- Returns JSON, fasta or plain text sequence
- Working with ENA to create a proposal for a Reference Sequence API



/seq/region/GRCh38/?location=12:43768112-43768272,43771220-43771365,43772180-43772362,43773045-43773072&translate=1

TaRK **Beta**

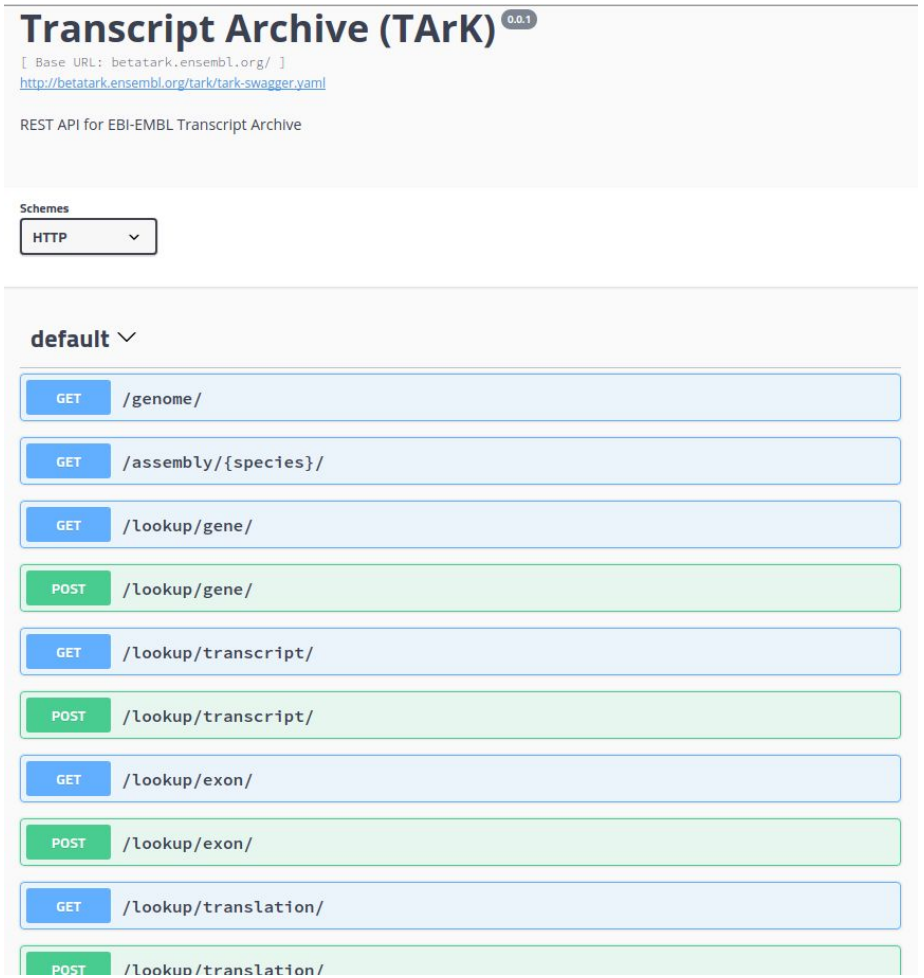


betatark.ensembl.org

TaRK **Beta**

- Swagger based site for exploring the API
- e86-90 loaded into the site
- **Beta** software, API may change at any time over the coming months
- Will become part of the Ensembl website backend

betatark.ensembl.org



Acknowledgements

Ensembl

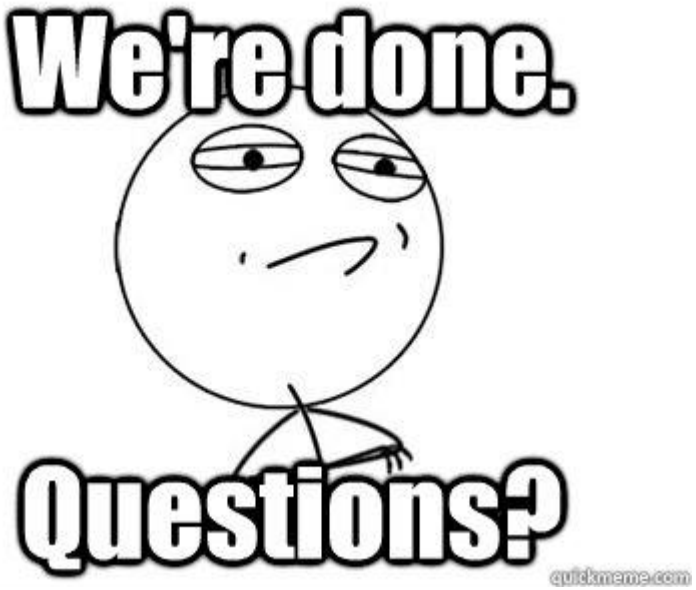
Premanand Achuthan
Magali Ruffier
Kieron Taylor
Alessandro Vullo
Daniel Zerbino

Bronwen Aken
Fiona Cunningham
Nick Langridge
Zhicheng Liu
Fergal Martin
Daniel Murphy
Andy Yates

Institute of Cancer Research

Nazneen Rahman
Elise Ruark
Ann Strydom

Global Alliance
for Genomics & Health


We're done.
Questions?
quickmeme.com