

Ensembl: How far will Elastic stretch?

Dan Staines
Genomics Technology Infrastructure

www.ensembl.org

www.ensemblgenomes.org



Data in Ensembl

- We offer...
 - 10,000s of genomes from 1000s of species
 - Gene model annotation or third-party import
 - Comparative genomics e.g. gene trees, alignments
 - Variation annotation
 - Regulatory elements
 - Cross-references, ontologies and protein features

Ensembl as a platform

- We provide...
 - A genome browser
 - Six sites for different parts of the taxonomy
 - MySQL databases
 - FTP flatfiles
 - Perl API
 - REST API

Ensembl: The browser



Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Human (GRCh38.p5) Location: 17:63,973,115-64,437,414

Login/Register

Search Human...

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- Region in detail**
- Comparative Genomics
 - Alignments (image)
 - Alignments (text)
 - Region Comparison
 - Synteny
- Genetic Variation
 - Resequencing
 - Linkage Data
- Markers
- Other genome browsers
 - UCSC
 - NCBI
 - Vega
 - Ensembl GRCh37

Configure this page

Add your data

Export data

Share this page

Bookmark this page

Chromosome 17: 63,973,115-64,437,414

Assembly exceptions

Chr. 17

Assembly exceptions

Region in detail

Chromosome bands

Contigs

Genes (Comprehensive set from GENCODE 24)

Gene Legend

- Ensembl protein coding
- processed transcript
- RNA gene
- merged Ensembl/Havana
- pseudogene

Location: 17:63973115-64437414 Go Gene: Go

Region in detail

Chromosome bands

Contigs

Human cDNAs (RefSeq/ENA)

Genes (Comprehensive)

AC005803.1 >

AC025362.12 >

AC016489.18 >

AC234063.4 >

ICAM2-006 protein coding

ICAM2-007 nonsense mediated decay

ICAM2-008 protein coding

ERN1-001 protein coding

ERN1-007 retained intron

ERN1-004 retained intron

ERN1-002 retained intron

ERN1-008 protein coding

TEX2-002 protein coding

TEX2-005 retained intron

TEX2-008 protein coding

TEX2-006 protein coding

TEX2-007 protein coding

RPL31P57-001 processed pseudogene

PECAM1-001 protein coding

PECAM1-009 processed transcript

Y_RNA.658-201 misc RNA

MIR1273E-201 miRNA

Beyond the browser: user needs

- Bulk export
 - e.g. names and descriptions on all genes from chr 1
- ID conversion
 - e.g. Ensembl IDs for a set of UniProt accessions
- Support for taxonomies and ontologies
 - e.g. all MAPK genes from Ascomycota
- Structured queries
 - e.g. genes annotated as involved in virulence in barley

Beyond the browser: BioMart

[New](#) [Count](#) [Results](#)
[URL](#) [XML](#) [Perl](#) [Help](#)

Dataset
 Homo sapiens genes (GRCh38.p2)

Filters
 Chromosome: 1

Attributes
 Ensembl Gene ID
 Ensembl Transcript ID
 HGNC ID(s)
 HGNC symbol


Dataset
 [None Selected]

Export all results to ☒ TSV ☐ Unique results only [Go](#)

Email notification to

View rows as ☒ HTML ☐ Unique results only

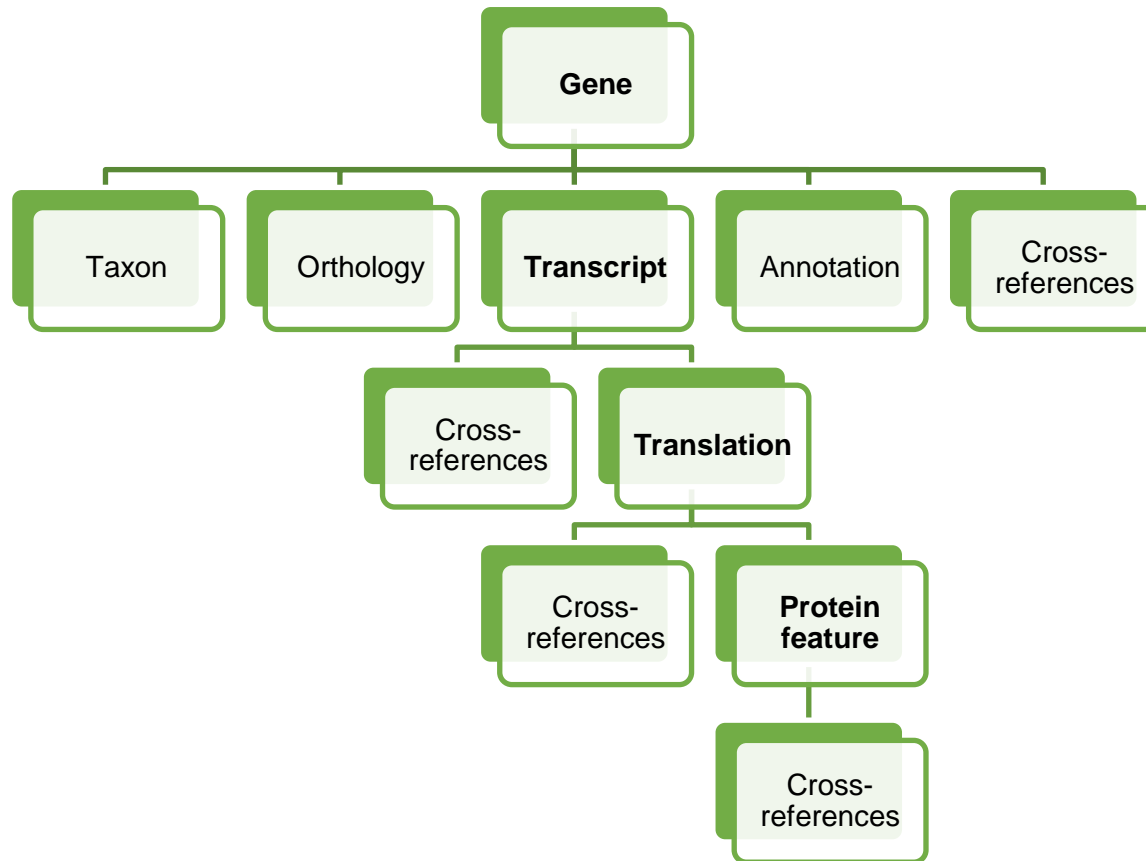
Ensembl Gene ID	Ensembl Transcript ID	HGNC ID(s)	HGNC symbol
ENSG00000142606	ENST00000471840	HGNC:14668	MMEL1
ENSG00000142606	ENST00000378412	HGNC:14668	MMEL1
ENSG00000142606	ENST00000502556	HGNC:14668	MMEL1
ENSG00000142606	ENST00000504800	HGNC:14668	MMEL1
ENSG00000142606	ENST00000491941	HGNC:14668	MMEL1
ENSG00000142606	ENST00000484195	HGNC:14668	MMEL1
ENSG00000142606	ENST00000469962	HGNC:14668	MMEL1
ENSG00000142606	ENST00000509374	HGNC:14668	MMEL1
ENSG00000142606	ENST00000511099	HGNC:14668	MMEL1
ENSG00000228750	ENST00000432429		



Beyond the browser: what next?

- Gene-centric searching
 - >100 million genes from >30,000 genomes
- Retain data structure
 - Support for nested queries
- Web interfaces to support wide variety of use-cases
- Programmatic APIs to support different languages e.g. R, Python, Java
- Use of search service in project-specific interfaces e.g. EBiSC, HipSci

Beyond the browser: what next?



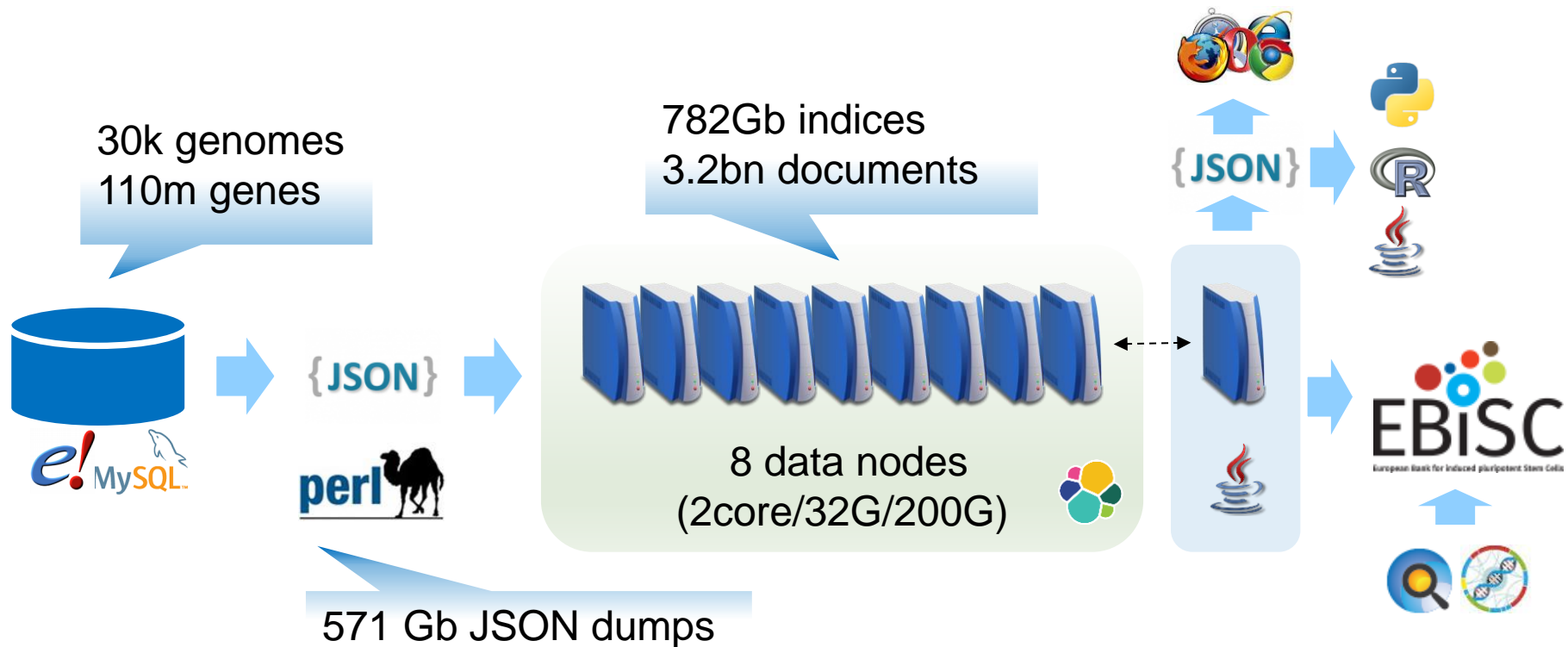
Beyond the browser: what next?



Beyond the browser: what next?



Elasticsearch: Taking it for a spin



Elasticsearch experiences: The best...

- ✓ Very easy to get started
- ✓ Excellent documentation
- ✓ Strong community
- ✓ Good support for nested data & aggregations
- ✓ Decent query and retrieval performance
- ✓ Comprehensive REST APIs
- ✓ Fast bulk loading via Perl API
- ✓ Nice ecosystem of free and commercial plugins
 - ✓ e.g. Marvel, elasticsearch-head

Elasticsearch experiences: The rest...

- Resource hungry
 - Cluster sizing feels like a dark art
 - Node failures can be very hard to recover from
- Some *very* unexpected behaviour acts as a roadblock
 - 100x leap in index size for some docs 2.0->2.1 (maybe norms?)
 - Unexplained spikes in heap size on some VMs (FS related?)
- No support for ontologies and taxonomies
- Parent-child type relationships aren't very performant
- Marvel & co are not free

Thanks to...



- Team GTI
 - Andy Yates
 - Kieron Taylor
- EBI Systems
 - Rich Boyce



The EBiSC - European Bank for induced pluripotent Stem Cells project has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115582, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. www.imi.europa.eu

